

François LIRET

Maths en pratique

À l'usage des étudiants

Cours et exercices



DUNOD

MATHS EN PRATIQUE

À l'usage des étudiants

MATHS EN PRATIQUE

À l’usage des étudiants

Cours et exercices

François Liret
Maître de conférence
à l’université Paris 7 – Denis Diderot

DUNOD

DU MÊME AUTEUR

Algèbre – Licence 1re année
(avec D. Martinais), Dunod, 2003

Analyse – Licence 1re année
(avec D. Martinais), Dunod, 2003

Algèbre et géométrie – Licence 2e année
(avec D. Martinais), Dunod, 2003

Analyse – Licence 2e année
(avec D. Martinais), Dunod, 2004

Illustration de couverture : Digital Vision



ISBN 2 10 049629 8

Table des matières

Chapitre 9. Limites, dérivées, intégrales

Chapitre 10. Utilisation de la dérivée et de l'intégrale

Chapitre 11. Interpolation, calcul numérique d'intégrales

Chapitre 12. Fonctions de plusieurs variables

Chapitre 13. Intégrales multiples

Chapitre 14. Champ de vecteurs, formes différentielles

Chapitre 15. Équations différentielles

Chapitre 16. Systèmes différentiels

Chapitre 17. Séries, séries entières, séries de Fourier

Avant-Propos

Ce livre s'adresse à des étudiants scientifiques d'un cursus Licence utilisant les Mathématiques comme outil de calcul, que ce soit dans le cadre de sciences générales pour l'ingénieur ou dans des disciplines spécifiques. C'est pourquoi l'on y présente de nombreux exemples d'applications dans des domaines variés, comme la Physique, les Sciences de la Vie, l'Économie ou la théorie du contrôle.

Les étudiants en Mathématiques, en Mathématiques appliquées ou en Informatique pourront aussi y découvrir des techniques de calcul, des exemples de modélisation et des problématiques que leur programme théorique ne laisse pas le temps d'explorer suffisamment.

Le livre se partage à peu près équitablement entre l'algèbre et l'analyse. S'agissant des méthodes numériques et des algorithmes présentés, le choix, nécessairement très sévère, s'est porté sur les plus courants et tient compte de l'efficacité, de la généralité et de la simplicité de mise en œuvre.

Dans le cours, les notions acquises dans une classe de Terminale scientifique sont supposées connues. Afin de développer des applications suffisamment riches tout en restant à un niveau élémentaire, de nombreux résultats ont été admis, parfois avec un commentaire explicatif ou heuristique appelé *justification*. Le terme *démonstration* est réservé à une argumentation complète sur le plan mathématique.

Les exercices sont des applications utiles et directes du cours. Il sont très généralement calculatoires, de sorte qu'on doit les faire à l'aide d'une calculatrice ou d'un logiciel de calcul scientifique. Quelques rares exercices apportent un petit complément théorique qu'il faut considérer comme un résultat à connaître. Le signe **@** indique que l'on trouvera une solution ou des indications dans les « Compléments en ligne » à l'adresse `http://www.dunod.com`.

Je remercie mes collègues de différentes disciplines qui ont bien voulu m'éclairer en répondant à mes questions et notamment Jacqueline concernant la problématique des Statistiques et des Probabilités. Merci à Eric pour son soutien technique et ses idées originales et à Michèle pour la marmotte.

Merci à Christian qui a relu le texte avec acuité : ses remarques pertinentes sont à l'origine de nombreuses améliorations. Un grand merci à Alberto qui, avec talent et gentillesse, a assuré un gros travail de mise en page et la réalisation finale de toutes les figures.

Et surtout, merci à Dominique.

Chapitre 1

Ensembles, nombres et fonctions

1. Langage et notations pour utiliser les ensembles

En Mathématiques, on définit souvent des collections d'objets appelées *ensembles*. Voici les notions générales et les expressions couramment employées lorsqu'on considère des ensembles.

1.1 Éléments et parties d'un ensemble

Éléments d'un ensemble. En général, on désigne un ensemble par une lettre. Si par exemple E désigne un ensemble, alors chaque objet a de la collection E s'appelle un *élément* de E : on dit que a *appartient à* E et l'on exprime cette propriété en écrivant $a \in E$. Le signe $\in$ est le symbole d'appartenance. Des ensembles sont égaux s'ils ont les mêmes éléments.

Parties d'un ensemble. Supposons que E est un ensemble. On appelle *partie de* E un ensemble formé de certains éléments de E. Une partie de E est donc un ensemble A ayant la propriété suivante : tout élément de A est aussi un élément de E. Pour exprimer que l'ensemble A est une partie de l'ensemble E, on dit aussi que A *est inclus dans* E, ce que l'on note $A \subset E$. Le signe d'inclusion $\subset$ ne s'écrit qu'entre deux ensembles et ne doit pas être confondu avec le signe d'appartenance $\in$.
Pour montrer que des ensembles E et F sont égaux, il faut vérifier que l'on a les deux inclusions $E \subset F$ et $F \subset E$.

Propriété caractéristique d'une partie. Supposons que l'on ait défini pour chaque élément x de E une propriété $\mathscr{P}(x)$ qui peut être satisfaite ou non. L'ensemble des éléments $x \in E$ qui satisfont la propriété $\mathscr{P}(x)$ est une partie de E que l'on note $\{x \in E \mid \mathscr{P}(x)\}$, ce qui se lit "l'ensemble des éléments x appartenant à E tels que $\mathscr{P}(x)$". Si l'on pose $A = \{x \in E \mid \mathscr{P}(x)\}$, la propriété $\mathscr{P}$ s'appelle la propriété caractéristique de A.

Exemple 1. Notons E l'ensemble des nombres entiers compris entre $-5/2$ et $11/3$. Les nombres 2 et 3 appartiennent à l'ensemble E, alors que -3 n'y appartient pas : on a donc les relations $2 \in E$ et $-3 \notin E$. Les éléments de l'ensemble E sont exactement les nombres -2, -1, 0, 1, 2 et 3 : pour exprimer cela, on écrit la liste des éléments entre accolades, sous la forme $E = \{-2, -1, 0, 1, 2, 3\}$.

Exemple 2. Un nombre entier positif ou nul s'appelle un *entier naturel* et l'ensemble de tous les entiers naturels se note $\mathbb{N}$. Par exemple, $168/6 \in \mathbb{N}$, $168/9$ n'est pas un entier naturel et pour tout entier $n \geqslant 1$, le nombre $(1+\sqrt{5})^n + (1-\sqrt{5})^n$ appartient à $\mathbb{N}$.

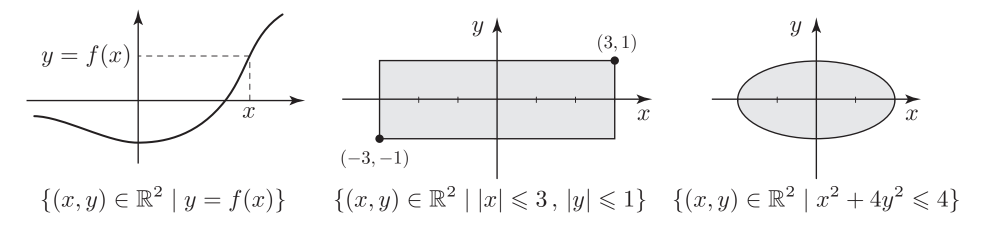
Exemple 3. Un nombre entier positif ou négatif ou nul, s'appelle un *entier relatif* et l'ensemble des entiers relatifs se note $\mathbb{Z}$. On a donc $\mathbb{N} = \{x \in \mathbb{Z} \mid x \geqslant 0\}$ et $\{x \in \mathbb{Z} \mid 0 < x^2 < 16\} = \{-3, -2, -1, 1, 2, 3\}$. L'ensemble E de l'exemple 1 est une partie de $\mathbb{Z}$ et l'on a $E = \{x \in \mathbb{Z} \mid 0 \leqslant x + 2 \leqslant 5\}$.

Définition

Soient E et F des ensembles. La donnée d'un élément x appartenant à E et d'un élément y appartenant à F s'appelle un *couple* et se note (x, y) ; la règle d'égalité est : $(x, y) = (x', y')$ si et seulement si $x = x'$ et $y = y'$. L'ensemble de ces couples se note $E \times F$ et s'appelle le *produit cartésien* des ensembles E et F.

Exemples 4. On utilise souvent des couples de nombres.

- Les couples $(1, 2)$ et $(2, 1)$ sont deux éléments différents appartenant à l'ensemble $\mathbb{N} \times \mathbb{N}$.
- L'ensemble des nombres réels se note $\mathbb{R}$. Les couples (x, y) de nombres réels sont les éléments de l'ensemble $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. Dans un plan muni d'axes, tout point est repéré par le couple (x, y) de ses coordonnées.
  On définit de même les triplets (x, y, z) de nombres réels : leur ensemble se note $\mathbb{R}^3$. Si l'on se donne un repère de l'espace, chaque point possède trois coordonnées : les triplets de nombres réels permettent de repérer les points de l'espace.
- Voici des ensembles de couples de nombres réels : la figure de gauche représente le graphe d'une fonction f, celle de droite l'intérieur d'une ellipse (voir page 22).



$\{(x, y) \in \mathbb{R}^2 \mid y = f(x)\}$ $\{(x, y) \in \mathbb{R}^2 \mid |x| \leqslant 3\,,\ |y| \leqslant 1\}$ $\{(x, y) \in \mathbb{R}^2 \mid x^2 + 4y^2 \leqslant 4\}$

1.2 Opérations sur les parties d'un ensemble

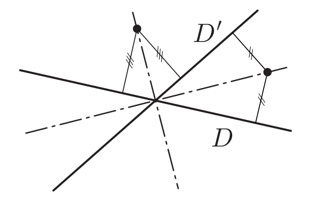
Intersection de parties. Soient A et B des parties d'un ensemble E. Les éléments de E qui appartiennent à la fois à A et à B forment une partie de E appelée *intersection de* A *et* B et notée $A \cap B$. On a donc $A \cap B = \{x \in E \mid x \in A \text{ et } x \in B\}$ et aussi les relations d'inclusion $A \cap B \subset A$ et $A \cap B \subset B$.
Si aucun élément de E n'appartient à l'intersection de A et B, on dit que la partie $A \cap B$ est vide et l'on écrit $A \cap B = \varnothing$. Le symbole $\varnothing$ désigne l'ensemble qui n'a aucun élément.

Réunion de parties. A et B étant des parties de E, l'ensemble des éléments de E qui appartiennent à A ou à B est une partie de E appelée *réunion de* A *et* B et notée $A \cup B$. On a les inclusions $A \subset A \cup B$ et $B \subset A \cup B$.

Complémentaire d'une partie. Si A est une partie de E, l'ensemble des éléments de E qui n'appartiennent pas à A s'appelle le *complémentaire de* A et se note $E \setminus A$. On a les relations $A \cup (E \setminus A) = E$ et $A \cap (E \setminus A) = \varnothing$.

Exemple 5. Donnons-nous deux points distincts P et Q dans l'espace euclidien et considérons l'ensemble H des points de l'espace qui sont équidistants de P et Q : c'est le plan passant par le milieu I du segment PQ et perpendiculaire à la droite (PQ). Le plan H s'appelle le plan médiateur de PQ. Soit R un autre point de l'espace différent de P et de Q et soit H' le plan médiateur de QR. Si les points P, Q, R ne sont pas alignés, l'intersection des plans H et H' est une droite D perpendiculaire au plan (PQR). Cette droite coupe le plan (PQR) en un point équidistant de P, Q et R, c'est-à-dire au centre du cercle circonscrit au triangle PQR. Si les points P, Q, R sont alignés, les plans H et H' sont parallèles et $H \cap H'$ est l'ensemble vide.

Exemple 6. Dans un plan euclidien, donnons-nous deux droites sécantes D et D'. L'ensemble des points du plan équidistants de D et D' est la réunion des deux bissectrices de l'angle formé par D et D'.



2. Les nombres

On exprime souvent un nombre par son écriture décimale. Pour un nombre entier ou décimal, il n'y a qu'un nombre fini de chiffres, alors qu'un nombre réel possède en général une écriture décimale illimitée.

2.1 Écriture d'un entier naturel

Les chiffres d'un entier naturel représentent les unités dans l'échelle $1, 10, 10^2, \ldots$ des puissances positives de 10 : par exemple, pour l'entier $a = 234$, on a $a = 4 + 3 \times 10 + 2 \times 10^2$.

Voici comment retrouver les chiffres d'un entier par des opérations arithmétiques.

- Puisque $a = 4 + 23 \times 10$, le chiffre des unités est le reste de la division de a par 10.
- L'entier $a - 4$ étant multiple de 10, posons $a - 4 = 10a_1$, donc $a_1 = (a-4)/10 = 23 = 3 + 2 \times 10$; le chiffre des dizaines est le reste de la division de a_1 par 10.
- L'entier $a_1 - 3$ est multiple de 10 et en posant $a_1 - 3 = 10a_2$, il vient $a_2 = (a_1 - 3)/10 = 2$ qui est le chiffre des centaines.

Pour écrire un algorithme de calcul des chiffres d'un entier $a > 0$, introduisons une variable x qui prendra successivement les valeurs $a, a_1, a_2, \ldots$ et une variable c dont les valeurs seront les chiffres de a ; la liste L des chiffres sera composée en écrivant de droite à gauche.

Algorithme de calcul des chiffres

initialisation : ($L \leftarrow$ liste vide) ($a \leftarrow$ un entier positif) ($x \leftarrow a$) ($b \leftarrow 10$)
boucle : tant que $x \neq 0$:
  $c \leftarrow$ reste de la division de x par b
  $x \leftarrow$ quotient de $x - c$ par b
  $L \leftarrow c, L$ *(on ajoute le chiffre c en tête de la liste L)*

Si au lieu de 10, on choisit pour b un entier naturel quelconque supérieur ou égal à 2, cet algorithme calcule les chiffres de *l'écriture de l'entier a en base b*, c'est-à-dire les entiers $c_0, c_1, \ldots, c_n$ tels que

$$0 \leqslant c_i < b \quad \text{et} \quad a = c_0 + c_1 \times b + c_2 \times b^2 + \cdots + c_n \times b^n .$$

La base 2 est commode, car il n'y figure que les chiffres 0 et 1. On a par exemple

$$13 = 1 + 2^2 + 2^3 \quad \text{et} \quad 34 = 2 + 2^5$$

donc, en base 2, ces nombres s'écrivent : $13 = [1101]$ et $34 = [100010]$.

2.2 Les entiers relatifs

Un entier relatif est un entier positif ou négatif ou nul. Dans l'ensemble $\mathbb{Z}$ des entiers relatifs, on peut donc ajouter, soustraire et multiplier.

Division euclidienne. Une autre opération essentielle qu'on peut effectuer avec des entiers relatifs, c'est la division avec reste, encore appelée *division euclidienne*. Rappelons la définition de la division par un entier $b > 0$.

Écrivons dans l'ordre les multiples entiers de b :

$$\ldots,\ -2b,\ -b,\ 0,\ b,\ 2b,\ \ldots,\ qb,\ \ldots \quad \text{où } q \in \mathbb{Z}.$$

Il est clair que tout entier relatif a est compris entre deux multiples consécutifs de b ; précisément, il y a un unique entier $q \in \mathbb{Z}$ tel que

$$qb \leqslant a < (q+1)b = qb + b\,.$$

On a alors $a - qb \geqslant 0$ et $a - qb < b$, donc en posant $r = a - qb$, il vient

$$a = qb + r\,, \quad \text{où } 0 \leqslant r < b.$$

L'entier q s'appelle le *quotient* de la division euclidienne de a par b et l'entier r s'appelle le *reste* de la division. Le reste est toujours un entier positif ou nul et strictement inférieur au diviseur b. Pour que l'entier a soit multiple de b, il faut et il suffit que le reste de la division de a par b soit nul.

Exemple. Étant donné un entier $n > 0$, cherchons le reste r_n de la division de 3^n par 8. Pour les premières valeurs de n, la division de 3^n par 8 s'écrit :

si $n = 1$: $3 = 0 \times 8 + 3$, donc $r_1 = 3$;
si $n = 2$: $3^2 = 9 = 1 \times 8 + 1$, donc $r_2 = 1$;
si $n = 3$: $3^3 = 3 \times 3^2 = 3(1 \times 8 + 1) = 3 \times 8 + 3$, donc $r_3 = 3$.

Il semble que les restes prennent successivement les valeurs 3 et 1. Pour vérifier cela, il suffit de montrer que l'on a $r_{n+2} = r_n$ pour tout n.
La division euclidienne de 3^n par 8 s'écrit $3^n = 8q_n + r_n$, où q_n désigne le quotient. Multiplions cette égalité par 3^2 ; il vient

$$3^{n+2} = 3^2 3^n = 3^2(8q_n) + 3^2 r_n = 8(3^2 q_n) + 8r_n + r_n = 8(3^2 q_n + r_n) + r_n\,.$$

On en déduit que r_n est aussi le reste de la division de 3^{n+2} par 8, donc on a l'égalité $r_{n+2} = r_n$. Puisque $r_1 = 3$ et $r_2 = 1$, on en déduit que $r_n = 3$ si n est impair et que $r_n = 1$ si n est pair.

2.3 Les nombres décimaux

Un *nombre décimal* est le produit d'un entier relatif et d'une puissance de 10. Par exemple, $12 \times 10^{-3} = 0{,}012$, $1234 \times 10^{-1} = 123{,}4$ et $(-3) \times 10^2 = -300$ sont des nombres décimaux. Pour avoir les chiffres d'un nombre décimal $a10^p$, où a est un entier, il suffit de décaler les chiffres de a de p places vers la gauche si l'entier p est positif, vers la droite si p est négatif. L'écriture décimale d'un nombre décimal ne comporte donc qu'un nombre fini de chiffres ; ce sont les unités dans l'échelle des puissances positives ou négatives de 10. Ainsi par exemple

$$308705 \times 10^{-4} = 30{,}8705 = 3 \times 10 + 8 \times 10^{-1} + 7 \times 10^{-2} + 5 \times 10^{-4}\,.$$

Un nombre décimal positif est la somme de sa partie entière et d'un nombre décimal a tel que $0 \leqslant a < 1$, donc l'écriture décimale de a est de la forme $a = 0, c_1 c_2 \cdots c_n$. Puisque $10a = c_1, c_2 \cdots c_n$, la première décimale c_1 est la partie entière du nombre $a_1 = 10a$. Si l'on pose $a_2 = 10a - c_1$, alors c_2 est la partie entière de $10a_2$. On calcule ainsi de proche en proche les décimales selon l'algorithme suivant :

Algorithme de calcul des décimales

initialisation : ($L \leftarrow$ liste vide) ($a \leftarrow$ un nombre décimal tel que $0 < a < 1$) ($x \leftarrow a$)
boucle : tant que $x \neq 0$:
  $c \leftarrow$ partie entière de $10x$
  $x \leftarrow 10x - c$
  $L \leftarrow L, c$ *(on ajoute le chiffre c à la liste L)*

L'ordre sur les décimaux. Considérons des nombres décimaux $a = 0, c_1 \cdots c_p$ et $a' = 0, c'_1 \cdots c'_p$ compris strictement entre 0 et 1 (on a écrit le même nombre de décimales en ajoutant éventuellement des zéros). Supposons que a et a' diffèrent seulement à partir de la k-ième décimale, les décimales précédentes étant égales ; alors on a : $a < a'$ si et seulement si $c_k < c'_k$. Par exemple, on a

$$0{,}01 = 0{,}01000 < 0{,}12339 = 0{,}12339000 < 0{,}12339001 < 0{,}123392 < 0{,}123400\,.$$

Si deux nombres décimaux ont des parties entières différentes, le plus grand est celui qui a la plus grande partie entière.

Densité des nombres décimaux. Si a est un nombre décimal, alors pour tout entier naturel n, le nombre $a' = a + 10^{-n}$ est décimal. En choisissant n assez grand, l'écart $a' - a = 10^{-n}$ peut être rendu aussi petit que l'on veut. Il y a donc des nombres décimaux différents de a et aussi près qu'on veut de a.

2.4 Les nombres réels

Une définition des nombres réels

Un nombre réel est de manière naturelle une limite de nombres décimaux. Soit $a[1], a[2], \ldots, a[n], \ldots$ des nombres décimaux positifs ; leurs écritures décimales sont de la forme

$$\begin{aligned} a[1] &= e[1], c_1[1]c_2[1] \cdots c_p[1] \cdots \\ a[2] &= e[2], c_1[2]c_2[2] \cdots c_p[2] \cdots \\ \cdots & \qquad \cdots \\ a[n] &= e[n], c_1[n]c_2[n] \cdots c_p[n] \cdots \end{aligned}$$

L'entier naturel $e[n]$ est la partie entière de $a[n]$ et les chiffres $c_1[n], c_2[n], \ldots$ sont les décimales de $a[n]$; on a terminé chaque écriture par des points de suspension pour éviter de préciser le nombre de chiffres, mais chacun des nombres $a[n]$ n'a qu'un nombre fini de décimales non nulles.
Supposons que la suite des nombres $a[1], a[2], \ldots, a[n], \ldots$ est décroissante, c'est-à-dire que l'on a $a[n] \geqslant a[n+1] > 0$ pour tout n.
Les parties entières vérifient donc $e[1] \geqslant e[2] \geqslant \cdots \geqslant e[n] \cdots$; puisque les $e[n]$ sont des entiers naturels, la suite $e[1], e[2], \ldots, e[n], \ldots$ finit par stationner à une valeur e, c'est-à-dire que pour tout entier n supérieur ou égal un certain rang N_0, on a $e[n] = e$. Pour $n \geqslant N_0$, on a $e[n+1] = e[n]$ et $a[n] \geqslant a[n+1]$, donc les premières décimales vérifient

$c_1[n] \geqslant c_1[n+1]$ pour tout $n \geqslant N_0$. En prenant la première décimale de chaque terme, on obtient après le rang N_0 une suite décroissante d'entiers naturels : après un certain rang N_1, la suite des entiers $c_1[n]$ est donc stationnaire à une valeur c_1. Ainsi on a

$$\begin{aligned}
a[N_0] &= \boldsymbol{e}, c_1[N_0]c_2[N_0]\cdots c_p[N_0]\cdots \\
\vdots \quad & \quad \vdots \\
a[N_1] &= \boldsymbol{e}, \boldsymbol{c_1} c_2[N_1]\cdots c_p[N_1]\cdots \\
\vdots \quad & \quad \vdots \quad \vdots \\
a[n] &= \boldsymbol{e}, \boldsymbol{c_1} c_2[n]\cdots c_p[n]\cdots
\end{aligned}$$

autrement dit $a[n] = e, c_1c_2[n]\cdots c_p[n]\cdots$ pour tout $n \geqslant N_1$.

Pour $n \geqslant N_1$, l'inégalité $a[n] \geqslant a[n+1]$ implique de même, pour les deuxièmes décimales, que l'on a $c_2[n] \geqslant c_2[n+1]$; il y a donc un rang N_2 après lequel les deuxièmes décimales gardent la même valeur c_2.

D'une manière générale, pour tout entier naturel p, il existe un rang N_p après lequel les p premières décimales restent fixes : pour tout $n \geqslant N_p$, on a $a[n] = e, c_1c_2\cdots c_pc_{p+1}[n]\cdots$, où les chiffres $c_1, c_2, \ldots, c_p$ ne dépendent plus de n pourvu que $n \geqslant N_p$.

Pour tout entier $p \geqslant 1$, on définit ainsi un chiffre c_p et l'on est conduit à convenir que l'écriture $e, c_1c_2\cdots c_p\cdots$ représente un nombre c, bien que cette écriture puisse présenter une infinité de chiffres non nuls. Le nombre c n'est pas décimal en général, mais les nombres décimaux $a[n]$ approchent c de mieux en mieux : en effet, pour $n \geqslant N_p$, les deux nombres décimaux $a[n]$ et $e, c_1c_2\cdots c_p$ ont mêmes p premières décimales, donc l'écart $a[n] - c$ est moindre que 10^{-p} ; puisque le nombre 10^{-p} peut être rendu aussi petit qu'on veut en choisissant p assez grand, on dit que *la suite des nombres $a[n]$ a pour limite c*, ce qu'on écrit sous la forme

$$\lim \big(a[n]\big) = c$$

Par définition, les nombres ainsi obtenus comme limite d'une suite décroissante de nombres décimaux positifs sont les *nombres réels* positifs ou nuls. Un nombre réel possède en général une infinité de décimales non nulles : on dit que son écriture décimale est illimitée.

Exemple. Posons $a[1] = 2/10$ et définissons une suite en posant

$$a[n+1] = \big(a[n]\big)^2 + (1/10) \quad \text{pour tout } n \geqslant 1.$$

Le nombre $a[1]$ est décimal et comme la somme et le produit de deux décimaux sont décimaux, tous les nombres $a[n]$ sont décimaux.

Montrons que la suite des $a[n]$ est décroissante, c'est-à-dire que l'on a $a[n+1] < a[n]$ pour tout entier $n \geqslant 1$. Nous raisonnons par récurrence. Puisque $a[2] = (2/10)^2 + 1/10 = 0{,}14$, on a bien $a[2] < a[1]$. Supposons que n est un entier tel que $a[n+1] < a[n]$. Puisque les nombres $a[n]$ sont tous positifs, il vient $(a[n+1])^2 < (a[n])^2$ donc

$a[n+2] = (a[n+1])^2 + 1/10 < (a[n])^2 + 1/10 = a[n+1]$. D'après le principe de récurrence, on en déduit que l'inégalité $a[n+1] < a[n]$ est vraie quel que soit l'entier $n \geqslant 1$.

La liste ci-contre donne les premières valeurs approchées des nombres $a(n)$: la limite des $a[n]$ est le nombre réel c dont les premières décimales sont 0,112701665··· On peut calculer algébriquement le nombre c : en effet, faisons tendre n vers l'infini dans l'égalité $a[n+1] = (a[n])^2 + (1/10)$; puisque $a[n]$ tend vers c, $(a[n])^2$ tend vers c^2 et comme $a[n+1]$ tend vers c, on obtient

$$c = c^2 + 1/10$$

ou encore $c^2 - c + 1/10 = 0$. L'équation du second degré $x^2 - x + 1/10$ a pour discriminant $1-(4/10)=3/5$ et pour racines les nombres réels $(1/2)(1+\sqrt{3/5})$ et $(1/2)(1-\sqrt{3/5})$. Puisqu'on a les inégalités

$$0 < c < 1/2 < (1/2)(1+\sqrt{3/5})\,,$$

on en déduit $c = (1/2)(1-\sqrt{3/5}) = \frac{5-\sqrt{15}}{10}$. On peut montrer que $\sqrt{15}$ n'est pas un nombre décimal, par suite le nombre c n'est pas décimal.

$a[1] = 0,2$
$a[2] = \mathbf{0,1}440000000$
$a[3] = \mathbf{0,11}96000000$
$a[4] = \mathbf{0,11}43041600$
$a[5] = \mathbf{0,11}30654409$
$a[6] = \mathbf{0,1127}837939$
$a[7] = \mathbf{0,1127}201841$
$a[8] = \mathbf{0,11270}58399$
$a[9] = \mathbf{0,11270}26063$
$a[10] = \mathbf{0,112701}8774$
$a[11] = \mathbf{0,1127017}131$
$a[12] = \mathbf{0,11270167}60$
$a[13] = \mathbf{0,112701667}0$
$a[14] = \mathbf{0,112701665}0$
$a[15] = \mathbf{0,112701665}5$
$a[16] = \mathbf{0,112701665}4$

Propriétés des nombres réels

L'ensemble des nombres réels se note $\mathbb{R}$. Rappelons que l'on peut faire la somme, la différence et le produit de deux nombres réels ; on peut aussi diviser par un nombre réel non nul, ce qui a pour conséquence la règle :

si a et b sont des nombres réels tels que $ab = 0$, alors $a = 0$ ou $b = 0$.

L'ordre sur les nombres réels. La comparaison entre deux nombres réels se définit comme pour les nombres décimaux : par exemple, si des nombres réels $c = 0,c_1c_2\cdots c_{k-1}c_k\cdots$ et $c' = 0,c_1c_2\cdots c_{k-1}c'_k\cdots$ ne diffèrent qu'à partir de la k-ième décimale, alors : $c < c'$ si et seulement si $c_k < c'_k$. On a aussi la définition plus algébrique :

$$c < c' \text{ si et seulement si } c' - c > 0.$$

Si a est un nombre réel, il existe des nombres réels différents de a et aussi près qu'on veut de a, par exemple les nombres $a + 10^{-n}$ pour n assez grand. Rappelons quelques définitions :

Définitions

- La *partie entière* d'un nombre réel a est l'entier (positif ou négatif) noté $E(a)$, tel que $E(a) \leqslant a < E(a) + 1$.
- Soient a et b des nombres réels tels que $a < b$. L'*intervalle ouvert* $]a,b[$ est l'ensemble des nombres réels x tels que $a < x < b$. L'*intervalle fermé* $[a,b]$, appelé aussi *segment*, est l'ensemble des nombres réels x tels que $a \leqslant x \leqslant b$.

► Si a et b sont des nombres réels, on définit de manière analogue les intervalles $]a,b]$, $[a,b[$, $]a,+\infty[$, $[a,+\infty[$, $]-\infty,b[$, $]-\infty,b]$ et l'on pose $]-\infty,+\infty[=\mathbb{R}$.

Voici les principales règles de calcul sur les inégalités entre nombres réels :

$$a < b \text{ si et seulement si } a + x < b + x$$
$$\text{si } a < b \text{ et } a' \leqslant b', \text{ alors } a + a' < b + b'$$
$$\text{si } x > 0, \text{ alors on a : } a < b \text{ si et seulement si } ax < bx$$
$$\text{si } x < 0, \text{ alors on a : } a < b \text{ si et seulement si } bx < ax$$
$$\text{pour des nombres } a \text{ et } b \text{ de même signe, on a : } a < b \text{ si et seulement si } 1/b < 1/a.$$

2.5 Les nombres rationnels

On appelle *nombre rationnel* le résultat de la division d'un nombre entier par un (autre) nombre entier non nul. Tout nombre décimal est rationnel, mais le nombre rationnel $4/3 = 1{,}33\cdots3\cdots$ n'est pas décimal puisqu'il a une infinité de décimales non nulles. L'ensemble des nombres rationnels se note $\mathbb{Q}$. On a donc les inclusions d'ensembles

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}\,.$$

On a souvent besoin d'approcher un nombre réel positif par des nombres rationnels. Si l'écriture décimale d'un nombre réel positif a est $e, c_1c_2\cdots c_n\cdots$, alors en ne retenant que les k premières décimales, on obtient des nombres décimaux $a[k] = e, c_1c_2\cdots c_k$ vérifiant $a[k] \leqslant a < a[k] + 10^{-k}$: chaque nombre décimal $a[k]$ approche a à 10^{-k} près. Posons $u_k = 10^k(0, c_1c_2\cdots c_k) = c_1c_2\cdots c_k$. Le nombre u_k est un entier et l'on a $a[k] = e + 0, c_1c_2\cdots c_k = \dfrac{10^k e + u_k}{10^k}$. L'approximation décimale à 10^{-k} près est donc en général une fraction dont le dénominateur est 10^k.

Meilleure approximation d'un nombre réel par des rationnels

Décrivons, sans la justifier, une méthode pour trouver des nombres rationnels à petits dénominateurs qui approchent rapidement un nombre réel $a > 0$. On suppose a non rationnel.

Posons $a_0 = E(a)$. Puisque a n'est pas entier, le nombre $a - a_0$ n'est pas nul et l'on peut définir un nombre x_1 en posant $a = a_0 + \dfrac{1}{x_1}$. Puisqu'on a $0 < a - a_0 < 1$, il vient $x_1 > 1$. Le nombre x_1 n'est pas entier, sinon a serait rationnel comme somme d'un entier et d'un rationnel. En posant $a_1 = E(x_1)$, on a donc $0 < x_1 - a_1 < 1$ et l'on peut définir un nombre $x_2 > 1$ tel que $x_1 = a_1 + \dfrac{1}{x_2}$. On continue ainsi, selon l'algorithme :

$$\begin{aligned}
a &= a_0 + \frac{1}{x_1}\ , \quad \text{où} \quad a_0 = E(a)\\
x_1 &= a_1 + \frac{1}{x_2}\ , \quad \text{où} \quad a_1 = E(x_1)\\
\cdots & \qquad \cdots\\
x_n &= a_n + \frac{1}{x_{n+1}}\ , \quad \text{où} \quad a_n = E(x_n)
\end{aligned}$$

Les deux premières lignes donnent $a = a_0 + \dfrac{1}{x_1} = a_0 + \dfrac{1}{a_1 + \dfrac{1}{x_2}}$ et l'on obtient ensuite les expressions suivantes, appelées *développement de a en fractions continuées* :

$$a = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{x_3}}} \quad , \quad a = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{x_4}}}}$$

Les nombres a_i sont des entiers positifs et les x_i sont des nombres réels plus grand que 1. Tronquons ces expressions en posant

$$b_0 = a_0 \quad , \quad b_1 = a_0 + \frac{1}{a_1} \quad , \quad b_2 = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2}} \quad , \quad \cdots$$

Les nombres $b_0, b_1, b_2, \ldots$ sont rationnels, car la somme et le produit de deux rationnels est rationnel et l'inverse d'un rationnel non nul est rationnel. On démontre que la suite des nombres b_n a pour limite a. Posons $b_n = \dfrac{p_n}{q_n}$, où p_n et q_n sont des entiers. Ces fractions ont une propriété remarquable : parmi les fractions $\dfrac{p}{q}$ telles que $q \leqslant q_n$, la fraction $b_n = \dfrac{p_n}{q_n}$ est la plus proche de a. En ce sens, les rationnels $\dfrac{p_n}{q_n}$ constituent la *meilleure approximation du nombre a par des fractions*. Ce qualificatif exprime que l'approximation est bonne bien que le dénominateur soit petit.

Exemple 1. Pour le nombre π, on a les égalités :

$$\pi = 3 + \frac{1}{x_1} \ , \qquad \text{où } x_1 = 7{,}0625133059310457697930051525 5\cdots$$

$$x_1 = 7 + \frac{1}{x_2} \ , \qquad \text{où } x_2 = 15{,}996594406685719888923060410 0\cdots$$

$$x_2 = 15 + \frac{1}{x_3} \ , \qquad \text{où } x_3 = 1{,}003417231013372603464147170 01\cdots$$

$$x_3 = 1 + \frac{1}{x_4} \ , \qquad \text{où } x_4 = 292{,}634591014395472378544147738\cdots$$

d'où les meilleures approximations rationnelles de π :

$b_0 = 3$

$b_1 = 3 + \dfrac{1}{7} = \dfrac{22}{7} \simeq 3{,}142\cdots$ (les deux premières décimales sont celles de π)

$b_2 = 3 + \cfrac{1}{7 + \cfrac{1}{15}} = 3 + \dfrac{15}{106} = \dfrac{333}{106} \simeq 3{,}14150\cdots$ (quatre décimales exactes)

$b_3 = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = 3 + \dfrac{16}{113} = \dfrac{355}{113} \simeq 3{,}1415929\cdots$ (six décimales exactes).

Exemple 2. En acoustique, on appelle intervalle entre deux sons le rapport de leurs fréquences : si des fréquences f_1, f_1', f_2, f_2' vérifient $f_1/f_1' = f_2/f_2'$, alors entre les sons de fréquences f_1 et f_1', l'oreille perçoit le même intervalle qu'entre les sons de fréquences f_2 et f_2'. Une note de musique de fréquence f s'accompagne d'harmoniques de fréquences $2f, 3f, \ldots$ et les notes de fréquences f et $2f$ sont perçues comme "les mêmes", jouées avec un écart d'une octave. L'intervalle entre les fréquences $2f$ et $3f$ s'appelle une quinte. Pour créer une gamme, c'est-à-dire pour découper les octaves en intervalles dont les multiples permettent suffisamment d'accords avec les quintes, on est amené à chercher de petits entiers naturels p et q tels que p quintes valent à peu près q octaves. Cette condition signifie que $(3/2)^p$ est peu différent de 2^q, ou encore que le rapport $2^{p+q}/3^p$ est proche de 1. En prenant le logarithme, cela se traduit par : $(p+q)\ln 2$ peu différent de $p \ln 3$, ou encore $\dfrac{\ln 3}{\ln 2}$ peu différent de la fraction $\dfrac{p+q}{p}$.
Cherchons donc les premières bonnes approximations rationnelles du nombre $\dfrac{\ln 3}{\ln 2}$.

$$\begin{aligned}
\frac{\ln 3}{\ln 2} &= 1{,}5849625007211561814537389439\cdots = 1 + (1/x_1)\\
x_1 &= 1{,}7095112913514547769761902621\cdots = 1 + (1/x_2)\\
x_2 &= 1{,}4094208396532090045824043308\cdots = 1 + (1/x_3)\\
x_3 &= 2{,}4424745961808592754871740323\cdots = 2 + (1/x_4)\\
x_4 &= 2{,}2600167526708245359312761226\cdots = 2 + (1/x_5)\\
x_5 &= 3{,}8459060415463995352281970839\cdots = 3 + (1/x_6)\ , \quad \text{où } x_6 > 1
\end{aligned}$$

d'où les approximations suivantes de $\dfrac{\ln 3}{\ln 2}$:

$$b_3 = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2}}} = \frac{8}{5} \quad , \quad b_4 = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{2}}}} = \frac{19}{12} \quad \text{et} \quad b_5 = \frac{65}{41}.$$

Avec l'approximation $\dfrac{8}{5}$, on obtient une gamme où cinq quintes sont équivalentes à trois octaves. Si f_0 est la fréquence de base d'une octave, on montre que les notes successives ont pour fréquence f_0, $f_1 = \dfrac{9}{8}f_0$, $f_2 = \dfrac{4}{3}f_0$, $f_3 = \dfrac{3}{2}f_0$, $f_4 = \dfrac{16}{9}f_0$: il y a donc cinq notes par octave. Entre deux sons successifs, il n'y a que deux intervalles possibles, car $f_1/f_0 = f_3/f_2 = 2f_0/f_4 = 9/8$ et $f_2/f_1 = f_4/f_3 = 32/27$; ces notes correspondent à peu près aux touches noires du piano.
Avec l'approximation $\dfrac{19}{12}$, on obtient une gamme chromatique, plus riche, où douze quintes valent sept octaves ; elle contient douze notes.

3. Les fonctions

3.1 Notion générale de fonction

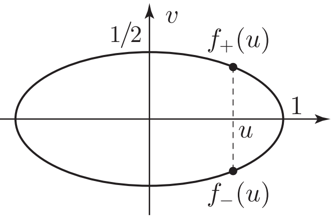
Soient E et F des ensembles. Une *application* ou une *fonction de E dans F* est une règle qui permet d'associer à chaque élément $x \in E$ un élément parfaitement déterminé de F. Il est souvent utile de nommer la fonction : lorsqu'on a défini une fonction f de E dans F, on note $f(x)$ l'élément de F associé à x par la fonction et l'on dit que $f(x)$ est *l'image* de x par la fonction f. La fonction elle-même se note $f : E \to F$ ou encore $x \mapsto f(x)$; l'ensemble E s'appelle *l'ensemble de départ* ou le *domaine de définition* de f ; l'ensemble F est *l'ensemble d'arrivée* de f.

Égalité de deux fonctions. Des fonctions f et g sont égales si et seulement si elles ont le même ensemble de départ, le même ensemble d'arrivée et si l'on a $f(x) = g(x)$ pour tout x appartenant à l'ensemble de départ.

Fonction constante. Une fonction $f : E \to F$ est *constante* si l'on a $f(x) = f(y)$ pour tous éléments x et y appartenant à E ; il revient au même de dire qu'il existe un élément $b \in F$ tel que $f(x) = b$ quel que soit $x \in E$.

Exemples

1) Supposons que des variables réelles u et v sont liées par la relation $u^2 + 4v^2 = 1$. Si l'on calcule v en fonction de u, on obtient la formule $v = \pm\frac{1}{2}\sqrt{1-u^2}$, valable pour les valeurs de u telles que $-1 \leqslant u \leqslant 1$. Cela permet de définir deux fonctions exprimant v au moyen de u : la fonction $f_+ : [-1,1] \to \mathbb{R}$ définie par $f_+(u) = \frac{1}{2}\sqrt{1-u^2}$ et la fonction $f_- : [-1,1] \to \mathbb{R}$ définie par $f_-(u) = -\frac{1}{2}\sqrt{1-u^2}$. Mais la relation $u^2 + 4v^2 = 1$ ne définit pas v comme fonction de u, car à une valeur de u appartenant à $]-1,1[$ correspond deux valeurs opposées de v.



2) Notons P la pression, V le volume et T la température absolue d'une masse gazeuse donnée. En première approximation, le rapport $\frac{PV}{T}$ reste égal à une constante k. Cette relation permet de définir plusieurs fonctions.
   - Si le volume V reste constant, la pression est fonction de la température selon la loi $T \mapsto P(T)$ définie par $P(T) = (k/V)T$. Les quantités $P(T)$ et T sont alors proportionnelles : on dit que la fonction P est *linéaire*.
   - À température constante, la pression est fonction du volume : cela définit une fonction $p : V \mapsto \frac{kT}{V}$. Les quantités $p(V)$ et V sont inversement proportionnelles.
   - La température est fonction à la fois du volume et de la pression, ce qui définit la fonction de deux variables $T : (P,V) \mapsto \frac{PV}{k}$; les quantités P et V

sont des nombres réels positifs, donc la fonction T est définie sur l'ensemble $]0,+\infty[\times]0,+\infty[$ des couples de nombres réels positifs.

Définition

Soit $f:E\rightarrow F$ une application. Si A est une partie de l'ensemble de départ E, l'ensemble des éléments de F de la forme $f(x)$, où x parcourt la partie A, s'appelle *l'image* de la partie A et se note $f(A)$. L'image de l'ensemble de départ, c'est-à-dire l'ensemble $f(E)$ de tous les éléments de la forme $f(x)$, s'appelle *l'image de* f.

L'image d'une partie de l'ensemble de départ est une partie de l'ensemble d'arrivée. Il ne faut pas confondre

- l'image d'un élément x : c'est l'élément $f(x)$ appartenant à l'ensemble d'arrivée,
- l'image d'une partie A : c'est une partie de l'ensemble d'arrivée.

Exemples

- Si $f:\mathbb{R}\rightarrow\mathbb{R}$ est la fonction $x\mapsto x^2$, alors $f\big([2,3[\big)=[4,9[$, $f\big([-2,3]\big)=[0,9]$ et pour tout nombre $a\leqslant 0$, on a $f\big([a,+\infty[\big)=[0,+\infty[$ (figure 1). L'image de f est l'intervalle $[0,+\infty[$.
- L'image de la fonction sinus est le segment $[-1,1]$. Puisque sinus a pour période 2π, on obtient toute l'image en prenant seulement les valeurs $\sin x$ quand x parcourt un intervalle de longueur 2π : pour tout nombre réel a, on a donc $\sin\big([a,a+2\pi[\big)=[-1,1]$.
- La figure 2 montre le graphe de la fonction $u:\mathbb{R}\rightarrow\mathbb{R}$ qui à x associe x^3-3x : l'image par u de $[-1,1]$ est $[-2,2]$ et l'image de l'application u est $\mathbb{R}$.
- Définissons une fonction $g:\mathbb{R}^2\rightarrow\mathbb{R}^2$ en posant $g(x,y)=(x^2+1,e^y)$. Les nombres de la forme x^2+1 sont tous les nombres supérieurs ou égaux à 1 ; les nombres qui peuvent s'écrire e^y sont les nombres strictement positifs : l'image de g est donc la partie $[1,+\infty[\times]0,+\infty[$ de $\mathbb{R}^2$ (figure 3).

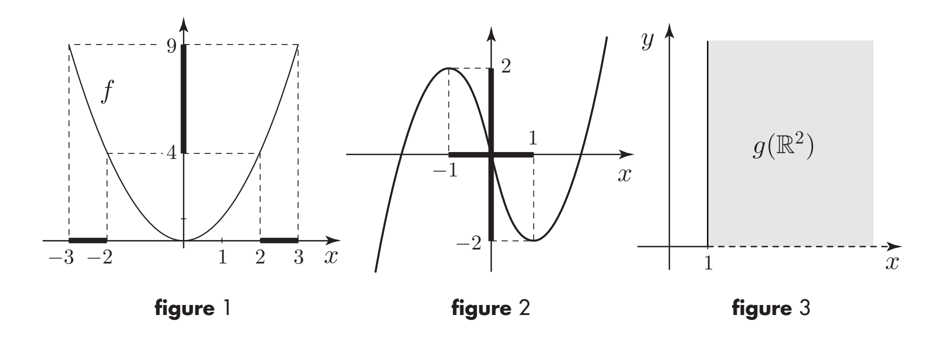


figure 1 **figure** 2 **figure** 3

Composée de deux fonctions. Soit $f : E \to F$ une fonction et soit $g : F \to G$ une fonction définie sur l'ensemble F d'arrivée de f. Pour tout $x \in E$, on a $f(x) \in F$, donc $g\big(f(x)\big)$ est un élément bien défini de l'ensemble G. L'association $x \mapsto g\big(f(x)\big)$ est donc une fonction de E dans G.

Définition

La fonction de E dans G qui à tout $x \in E$ associe l'élément $g\big(f(x)\big)$ s'appelle la *composée* des fonctions f et g et se note $g \circ f : E \to G$.

Si l'on a encore une fonction $h : G \to H$, les composées

$$h \circ (g \circ f) : \ E \xrightarrow{g \circ f} G \xrightarrow{h} H \qquad \text{et} \qquad (h \circ g) \circ f : \ E \xrightarrow{f} F \xrightarrow{h \circ g} H$$

sont des fonctions de E dans H. Comme on a l'égalité $(h \circ g) \circ f(x) = h \circ (g \circ f)(x)$ pour tout x, les fonctions $(h \circ g) \circ f$ et $h \circ (g \circ f)$ sont égales et l'on omet les parenthèses en écrivant simplement $h \circ g \circ f$ cette composée.

Exemples

1) Supposons qu'une quantité w dépende du temps et de la distance à un point O. Choisissons un repère orthonormé $(O; \vec{i}, \vec{j}, \vec{k})$ d'origine O. Cela définit une fonction $w : (x,y,z,t) \mapsto w(x,y,z,t)$, où t est la variable temps et x, y, z les coordonnées spatiales. Comme w ne dépend que de t et de la distance $r(x,y,z) = \sqrt{x^2+y^2+z^2}$ au point O, posons $f(r,t) = w(x,y,z,t)$: la fonction f n'a que deux variables et w est la composée $w = f \circ g$, avec $g(x,y,z,t) = (\sqrt{x^2+y^2+z^2}, t)$:

$$\begin{array}{ccc} \mathbb{R}^3 \times \mathbb{R} & \xrightarrow{g} & \mathbb{R} \times \mathbb{R} \\ (x,y,z,t) & \mapsto & (\sqrt{x^2+y^2+z^2}, t) \end{array} \qquad \begin{array}{ccc} \mathbb{R} \times \mathbb{R} & \xrightarrow{f} & \mathbb{R} \\ (r,t) & \mapsto & f(r,t) \end{array}$$

2) Posons $f(x,y) = \dfrac{xy}{x^2+y^2}$, où x et y sont des nombres réels strictement positifs : cela définit une fonction $f : E \to \mathbb{R}$, où $E =]0,+\infty[\times]0,+\infty[$. En divisant numérateur et dénominateur par y^2, il vient $f(x,y) = \dfrac{x/y}{(x/y)^2+1} = \dfrac{t}{t^2+1}$, où $t = x/y$. Ainsi les valeurs $f(x,y)$ ne dépendent que du rapport x/y. Si l'on définit la fonction $u :]0,+\infty[\to \mathbb{R}$ en posant $u(t) = \dfrac{t}{t^2+1}$, alors on a $f(x,y) = u(x/y)$, autrement dit f est la composée $f = u \circ g$, où $g : E \to]0,+\infty[$ est définie par $g(x,y) = x/y$. Quand les valeurs $\varphi(x,y)$ d'une fonction ne dépendent que du rapport x/y (comme c'est le cas pour f), on dit que la fonction φ est *homogène*.

3) Définissons des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ en posant $f(x) = x^2$ et $g(x) = x+1$: on a $g\big(f(x)\big) = x^2+1$ et $f\big(g(x)\big) = (x+1)^2$, donc $f \circ g \neq g \circ f$. Dans le cas où les composées $f \circ g$ et $g \circ f$ sont toutes les deux définies, on voit qu'en général, l'ordre de composition compte.

Rappelons les définitions relatives aux fonctions monotones.

Définitions

Soit f une fonction à valeurs réelles définie sur une partie de l'ensemble $\mathbb{R}$.

- f est *croissante* si, pour tous nombres x et y appartenant à l'ensemble de départ et vérifiant $x \leqslant y$, on a $f(x) \leqslant f(y)$;
- f *strictement croissante* si, pour tous nombres x et y appartenant à l'ensemble de départ et vérifiant $x < y$, on a $f(x) < f(y)$.
- On définit de même une fonction *décroissante* et une fonction *strictement décroissante*.

La composée de deux fonctions croissantes est croissante, de même que la composée de deux fonctions décroissantes. Si l'on compose une fonction croissante et une fonction décroissante, on obtient une fonction décroissante.

3.2 Transformation et itération

Une fonction $f : E \to E$ s'appelle une *transformation de* E.
Si f est une transformation de E, alors pour tout élément x_0 de E, l'élément $x_1 = f(x_0)$ appartient encore à E et l'on peut définir les éléments $x_2 = f(x_1)$, $x_3 = f(x_2)$, etc. On forme ainsi les *itérés* $x_1, x_2, \ldots, x_n, \ldots$ de x_0 par la transformation f : ils sont définis de proche en proche par la relation

$$\text{pour tout entier } n \geqslant 0\,, \quad x_{n+1} = f(x_n).$$

La transformation $x_0 \mapsto x_n$ est réalisée par la fonction composée $\underbrace{f \circ f \circ \cdots \circ f}_{n \text{ fois}}$, que l'on note $f^n : E \to E$.

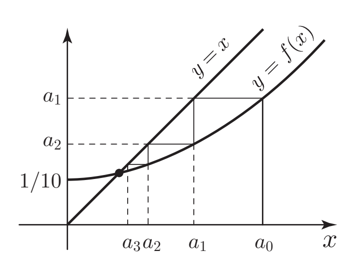
Définition

Si $f : E \to E$ est une transformation, un élément $x \in E$ tel que $f(x) = x$ s'appelle un *point fixe* de f.

Si x est un point fixe de f, tous les itérés de x sont égaux à x.

Exemple 1. Reprenons les nombres a_n définis (exemple page 7) par les relations $a_0 = 2/10$ et $a_{n+1} = a_n^2 + (1/10)$ pour tout entier $n \geqslant 0$. En posant $f(x) = x^2 + (1/10)$, on obtient une fonction $f :]0, +\infty[\to]0, +\infty[$ et les nombres a_n sont les itérés par f du nombre $2/10$.

Les points fixes de f sont par définition les nombres positifs solutions de l'équation $x^2 + (1/10) = x$. Puisque les solutions $(1/2)(1+\sqrt{3/5})$ et $(1/2)(1-\sqrt{3/5})$ sont des nombres positifs, ce sont les deux points fixes de f. Dans l'exemple, nous avions observé que les itérés de a_0 tendent vers l'un des points fixes de f.



La figure ci-contre montre la courbe d'équation $y = f(x)$

et la droite d'équation $y = x$ sur l'intervalle $[0{,}2/10]$. On y a construit une ligne brisée formée de segments alternativement horizontaux et verticaux, de la manière suivante :

- le segment vertical le plus à droite est à l'abscisse $a_0 = 2/10$, donc coupe le graphe de f au point d'ordonnée $f(a_0) = a_1$;
- le segment horizontal qui suit est à l'ordonnée a_1, valeur qu'on reporte sur l'axe des abscisses en passant par la bissectrice $y = x$: on a visiblement $a_1 < a_0$;
- le segment vertical d'abscisse a_1 coupe le graphe de f au point d'ordonnée $f(a_1) = a_2$; reportons cette valeur a_2 sur l'axe des abscisses comme on l'a fait pour a_1 : on observe que $a_2 < a_1$.

Cette construction permet de visualiser sur l'axe des x l'évolution des itérés de a_0 : la suite des nombres a_n est décroissante et sa limite est l'abscisse ℓ du point d'intersection de la courbe et de la bissectrice $y = x$. On a bien l'égalité $\ell = f(\ell)$ qui traduit que ℓ est un point fixe de f. Le seul point fixe de f inférieur à a_0 est $(1/2)(1-\sqrt{3/5})$, par conséquent $\ell = (1/2)(1-\sqrt{3/5})$.

Exemple 2. Soient a et b des nombres réels et soit $f : \mathbb{R} \to \mathbb{R}$ la fonction définie par $f(x) = ax + b$.
Si $a = 1$ et $b = 0$, on a $f(x) = x$ pour tout x, donc tous les nombres réels sont des points fixes de f. Si $a = 1$ et $b \neq 0$, on a $f(x) = x + b$ donc il n'y a aucun point fixe.
Supposons $a \neq 1$. Un nombre x est point fixe de f si et seulement si $ax + b = x$, ce qui équivaut à $(1 - a)x = b$. L'unique point fixe est donc $\omega = b/(1 - a)$.
On a $f(x) - f(\omega) = ax + b - (a\omega + b) = a(x - \omega)$ et puisque $f(\omega) = \omega$, il vient

$$f(x) - \omega = a(x - \omega) \quad \text{pour tout } x.$$

Donnons-nous un nombre réel x_0 et exprimons les itérés $x_1 = ax_0 + b$, $x_2 = ax_1 + b$, ... , $x_n = ax_{n-1} + b$ de x_0. On a $x_1 - \omega = a(x_0 - \omega)$, $x_2 - \omega = a(x_1 - \omega) = a^2(x_0 - \omega)$ et en général

$$x_n - \omega = a^n(x_0 - \omega) \quad \text{pour tout entier } n \geqslant 1,$$

formule qui s'écrit encore

$$x_n = a^n x_0 + \omega(1 - a^n) = a^n x_0 + b\frac{1 - a^n}{1 - a}.$$

Exemple 3. Dans des conditions stables, deux espèces A et B de bactéries vivent en symbiose à des concentrations moyennes a et b. On déplace l'équilibre en augmentant de 18% la concentration de A et de 12% celle de B, puis on mesure chaque jour l'écart à la moyenne des concentrations de chaque espèce. Au bout de n jours, l'écart pour la bactérie A, en pourcentage de a, vaut x_n et pour la bactérie B, il vaut y_n. Après modélisation, on est conduit à la loi d'évolution suivante :

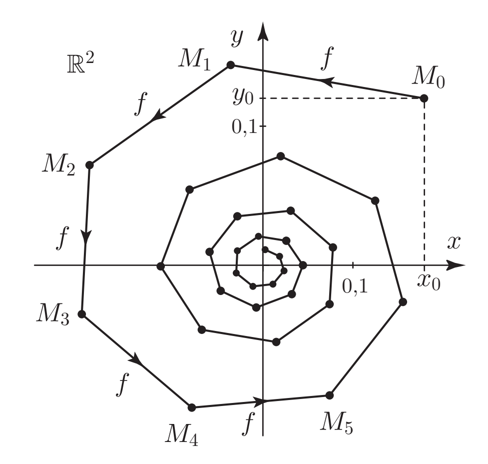
$$\begin{cases} x_{n+1} = (1/5)(3x_n - 6y_n) \\ y_{n+1} = (1/5)(2x_n + 3y_n) \end{cases}$$

avec $x_0 = 0{,}18$ et $y_0 = 0{,}12$ d'après nos conditions initiales.

Les couples (x_n, y_n) sont les itérés de (x_0, y_0) par la fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$f(x,y) = (X,Y) \ , \quad \text{où } X = \frac{1}{5}(3x - 6y) \text{ et } Y = \frac{1}{5}(2x + 3y)\,.$$

On vérifie facilement que le seul point fixe de f est $(0,0)$: cela veut dire que les concentrations de A et B ne peuvent rester en équilibre que pour les valeurs a et b.
Le graphique ci-contre montre les points $M_0{=}(x_0,y_0), M_1{=}(x_1,y_1), \ldots, M_{30}{=}(x_{30},y_{30})$: ils se situent sur une courbe en spirale qui entoure l'origine et s'en rapproche, x_n et y_n pouvant prendre des valeurs positives ou négatives. Observons la distance à l'origine d'un point M_n : elle n'est pas décroissante mais semble cependant tendre vers 0, ce qui indique que M_n tend vers l'origine quand n tend vers l'infini.



Pour mesurer l'éloignement à l'origine d'un point (x,y), nous allons utiliser une fonction $\delta(x,y) = x^2 + ey^2$, où $e > 0$. Si d est un nombre positif, les points (x,y) tels que $\delta(x,y) = d$, c'est-à-dire la courbe d'équation $x^2 + ey^2 = d$, est une ellipse centrée à l'origine (exemple 1 page 22). Les points (x,y) tels que $\delta(x,y) \leqslant d$ sont à l'intérieur de l'ellipse et plus d est petit, plus ces points sont proches de l'origine : $\delta(x,y)$ peut donc servir à mesurer l'éloignement à l'origine. Remarquons que si l'on choisit $e = 1$, $\delta(x,y)$ est le carré de la distance à l'origine et les ellipses sont des cercles.
Comparons les nombres $\delta(X,Y)$ et $\delta(x,y)$.

$$\delta(X,Y) - \delta(x,y) = X^2 + eY^2 - x^2 - ey^2 = \frac{1}{25}\left((3x-6y)^2 + e(2x+3y)^2 - 25x^2 - 25ey^2\right)$$
$$= \frac{1}{25}\left(4(e-4)x^2 + 12(-3+e)xy + 4(9-4e)y^2\right).$$

Choisissons $e = 3$. Il vient

$$\delta(X,Y) - \delta(x,y) = \frac{1}{25}\left(-4x^2 - 12y^2\right) = -\frac{4}{25}(x^2 + 3y^2) = -\frac{4}{25}\delta(x,y)\,,$$

d'où $\delta(X,Y) = \dfrac{21}{25}\delta(x,y) = 0{,}84\,\delta(x,y)$ et finalement

$$\delta(x_{n+1}, y_{n+1}) = (0{,}84)\,\delta(x_n, y_n) \quad \text{pour tout } n.$$

La suite des nombres $\delta(x_n, y_n)$ est géométrique, de premier terme $\delta(x_0, y_0) = \delta(0{,}18\,,\,0{,}12) = 0{,}0756$ et de raison $0{,}84$, de sorte que

$$\delta(x_n, y_n) = (0{,}84)^n\,\delta(x_0, y_0) \qquad \text{pour tout } n.$$

Puisque le nombre $0{,}84$ est strictement inférieur à 1, les puissances $(0{,}84)^n$ tendent vers 0 et $\delta(x_n, y_n)$ aussi. Remarquons que l'on a $|x| \leqslant \sqrt{x^2 + 3y^2}$ et $|y| \leqslant \sqrt{x^2 + 3y^2}$, donc $|x_n| \leqslant \sqrt{\delta(x_n, y_n)}$ et $|y_n| \leqslant \sqrt{\delta(x_n, y_n)}$. Il s'ensuit que les nombres x_n et y_n tendent vers 0.

Les concentrations des bactéries A et B reviennent donc vers les valeurs d'équilibre a et b en fluctuant autour de cet équilibre. On peut estimer la vitesse de retour vers l'équilibre en cherchant n pour que $|x_n|$ et $|y_n|$ soient, par exemple, inférieurs à 10^{-2} : il suffit pour cela que l'on ait $\delta(x_n, y_n) \leqslant 10^{-4}$, ou encore $(0{,}84)^n \times 0{,}0756 \leqslant 10^{-4}$. En prenant le logarithme, cette inégalité devient $n \ln(0{,}84) + \ln(0{,}0756) \leqslant -4\ln(10)$ c'est-à-dire $n \geqslant \dfrac{4\ln(10) + \ln(0{,}0756)}{-\ln(0{,}84)} = 38{,}01\cdots$, d'où $n \geqslant 39$.

Exemple 4. Dans un secteur de production, on diminue chaque année de 20% la pollution produite, mais parallèlement le coût énergétique présente une variation relative $0{,}25 - 0{,}6/r$, où r est le rapport coût sur pollution, mesuré à l'aide d'unités convenables. À une certaine date, on mesure une pollution annuelle p_0 et un coût correspondant c_0. Comment évolueront la pollution produite et le coût en énergie ?

Notons p_n la pollution produite pendant la n-ième année et c_n le coût énergétique. Le taux $\dfrac{p_{n+1} - p_n}{p_n}$ est égal à $-0{,}2$. Si l'on note r_n le rapport c_n/p_n, alors le taux $\dfrac{c_{n+1} - c_n}{c_n}$ est égal à $0{,}25 - 0{,}6/r_n$.

Exprimons le couple (p_{n+1}, c_{n+1}) au moyen de (p_n, c_n). L'égalité $\dfrac{p_{n+1} - p_n}{p_n} = -0{,}2$ conduit à $p_{n+1} = 0{,}8p_n$ et puisqu'on a $\dfrac{c_{n+1} - c_n}{c_n} = 0{,}25 - 0{,}6(p_n/c_n)$, il vient $c_{n+1} = 1{,}25c_n - 0{,}6p_n$. On a donc le système d'égalités

$$\begin{cases} p_{n+1} = 0{,}8p_n \\ c_{n+1} = 1{,}25c_n - 0{,}6p_n \end{cases}$$

Si l'on définit la fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$ en posant $f(p,c) = (0{,}8p\ ,\ 1{,}25c - 0{,}6p)$, ces égalités expriment que les couples (p_n, c_n) sont les itérés du couple (p_0, c_0) par la fonction f. D'après la première égalité, p_n suit une progression géométrique : on a donc

(1) $$p_n = (0{,}8)^n p_0\,.$$

En divisant membre à membre les égalités du système, il vient

$$r_{n+1} = \frac{c_{n+1}}{p_{n+1}} = \frac{1{,}25c_n}{0{,}8p_n} - \frac{0{,}6p_n}{0{,}8p_n} = \frac{12{,}5}{8} r_n - \frac{0{,}6}{0{,}8} = \frac{25}{16} r_n - \frac{3}{4}\,.$$

Les nombres r_n s'obtiennent en itérant la fonction $r \mapsto (25/16)r - (3/4)$; comme dans l'exemple 2, on a donc

$$r_n = (25/16)^n r_0 - (3/4)\frac{(25/16)^n - 1}{9/16} = (25/16)^n r_0 - (4/3)\big[(25/16)^n - 1\big]\,.$$

Multiplions par $p_n = (0{,}8)^n p_0$ en remarquant que l'on a $0{,}8 \times (25/16) = 0{,}05 \times 25 = 1{,}25$:

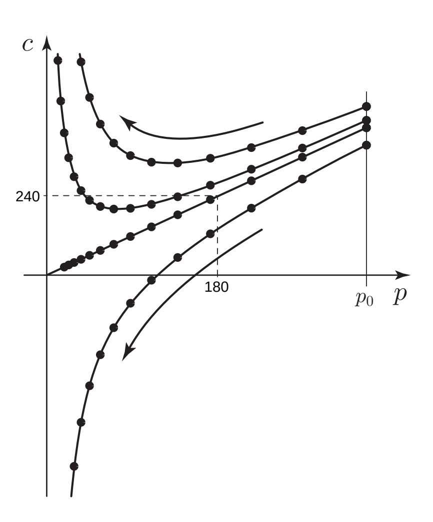
(2) $$c_n = (1{,}25)^n c_0 - (4/3)\big[(1{,}25)^n - (0{,}8)^n\big] p_0\,.$$

Sur la figure ci-après, nous avons représenté les points de coordonnées (p_n, c_n) pour les premières valeurs de n, avec différentes données initiales p_0 et c_0.

Pour calculer l'équation de la courbe où se trouvent les points (p_n, c_n), éliminons n entre les expressions données en (1) et (2). On a $(0{,}8)^n = p_n/p_0$ et comme $1{,}25 \times 0{,}8 = 1$, il vient $(1{,}25)^n = p_0/p_n$. D'où

$$c_n = \frac{p_0}{p_n} c_0 - \frac{4}{3}\left[\frac{p_0}{p_n} - \frac{p_n}{p_0}\right] p_0 = \frac{p_0}{p_n}\left[c_0 - \frac{4}{3} p_0\right] + \frac{4}{3} p_n$$

$$(3c_n - 4p_n)p_n = (3c_0 - 4p_0)p_0\,.$$



Ainsi la quantité $(3c_n - 4p_n)p_n$ reste constante au cours de l'itération, autrement dit les points (p_n, c_n) sont sur la courbe d'équation $(3c-4p)p = k$, où $k = (3c_0 - 4p_0)p_0$. En prenant p comme variable, il vient $c(p) = (1/3)(4p + k/p)$, la fonction c étant définie sur $]0, +\infty[$. La dérivée est $c'(p) = (1/3)(4 - k/p^2)$.

Premier cas : $3c_0 > 4p_0$. On a alors $k > 0$, la dérivée s'annule si $4p^2 = k$, c'est-à-dire si $p = \sqrt{k}/2$, d'où le tableau de variations :

p	$0 \qquad \sqrt{k}/2 \qquad +\infty$
$c(p)$	$\searrow \qquad \nearrow$

Puisque la pollution diminue chaque année, la courbe est parcourue dans le sens des abscisses décroissantes et la pollution tend vers 0. En supposant $p_0 > \sqrt{k}/2$, c'est-à-dire $4p_0 > 3c_0/2$, le coût commence par diminuer jusqu'à la valeur minimum $c(\sqrt{k}/2)$, puis augmente et tend vers l'infini. Si n est l'entier tel que $p_{n+1} \leqslant \sqrt{k}/2 < p_n$, le coût minimum est obtenu la $(n+1)$-ième année.

Deuxième cas : $3c_0 < 4p_0$. La constante k est négative, la dérivée $c'(p)$ est strictement positive et la fonction $p \mapsto c(p)$ est strictement croissante. Quand p tend vers 0, $c(p)$ tend vers $-\infty$ et quand p tend vers $+\infty$, $c(p)$ tend vers $+\infty$. Puisque la courbe est parcourue dans le sens des abscisses décroissantes, on atteint ainsi un point où $c = 0$, ce qui est irréaliste car le coût énergétique est une quantité positive.

Troisième cas : $3c_0 = 4p_0$. Les valeurs de c_0 et de p_0 étant le résultat de mesures, l'égalité $3c_0 = 4p_0$ n'est sûrement pas réalisée exactement : il s'agit d'un cas limite. D'un point de vue mathématique, on a alors $c(p) = (4/3)p$: la courbe est la demi-droite de pente 4/3 passant par l'origine. Le rapport c_n/p_n reste égal à 4/3 et les points de coordonnées (p_n, c_n) tendent vers l'origine quand n tend vers $+\infty$.

3.3 Notion d'antécédent, application bijective

Définition

Soit $f : E \to F$ une application. Si b est un élément de F, alors tout élément $x \in E$ tel que $f(x) = b$ s'appelle un *antécédent* de b par l'application f.

- Si b est un élément de l'ensemble d'arrivée, *résoudre l'équation* $f(x) = b$, c'est chercher tous les antécédents de b.
- Pour qu'un élément $b \in F$ ait au moins un antécédent par f, il faut et il suffit que b appartienne à l'image de f.

Exemples

1) Considérons la fonction valeur absolue $x \mapsto |x|$. Si b est un nombre réel strictement positif, alors l'équation $|x| = b$ possède deux solutions b et $-b$, donc b a deux antécédents. Un nombre strictement négatif n'a pas d'antécédent. Le nombre 0 a pour seul antécédent 0.

2) Donnons-nous un nombre réel b et étudions l'équation $\sin x = b$. Si $|b| > 1$, cette équation n'a pas de solution, car pour tout nombre réel x, on a $-1 \leqslant \sin x \leqslant 1$; un nombre de valeur absolue strictement supérieure à 1 n'a donc pas d'antécédent par la fonction sinus. Supposons $|b| \leqslant 1$; alors il existe un (unique) nombre $a \in [-\pi/2, \pi/2]$ tel que $\sin a = b$. On a les équivalences

$$\sin x = b \iff \sin x = \sin a \iff (x = a + 2k\pi \text{ ou } x = \pi - a + 2k\pi), \text{ où } k \in \mathbb{Z}.$$

Tout nombre $b \in [-1, 1]$ a donc une infinité d'antécédents par la fonction sinus.

3) Pour tout nombre réel $b \in [-1, 1]$, l'équation $\sin x = b$ possède une seule solution dans l'intervalle $[-\pi/2, \pi/2]$. Si l'on définit la fonction $s : [-\pi/2, \pi/2] \to [-1, 1]$ en posant $s(x) = \sin x$, alors tout élément de l'intervalle $[-1, 1]$ possède un unique antécédent par s.

4) **Un exemple arithmétique.** Si b est un entier donné, quelles sont les façons de l'écrire sous la forme $5x + 7y$ avec x et y des entiers relatifs ?
Définissons la fonction $g : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ en posant $g(x,y) = 5x + 7y$. Pour tout entier $b \in \mathbb{Z}$, on a $g(3b, -2b) = 5 \times 3b + 7 \times (-2b) = b$, donc le couple $(3b, -2b)$ est un antécédent de b. Cherchons les autres antécédents de b. Ce sont les couples (x,y) d'entiers vérifiant

$$5x + 7y = b \iff 5x + 7y = 5(3b) + 7(-2b) \iff 5(x - 3b) = -7(y + 2b).$$

L'entier $5(x - 3b)$ est multiple de 7 (car $y + 2b$ est entier). Puisque 7 et 5 sont premiers entre eux, on en déduit que l'entier $x - 3b$ est multiple de 7. On a donc $x - 3b = 7k$, où $k \in \mathbb{Z}$. Il vient $5(x - 3b) = 5 \times 7k = 7(-y - 2b)$, d'où $5k = -y - 2b$. Finalement, les antécédents de b sont les couples $(3b + 7k, -2b - 5k)$, où k est un entier relatif quelconque. Ce sont les couples (x, y) d'entiers tels que $b = 5x + 7y$.

Pour une fonction définie sur une partie de $\mathbb{R}$ et à valeurs réelles, l'étude du sens de variation renseigne sur le nombre d'antécédents d'un élément.

Proposition. *Soit $f : I \to \mathbb{R}$ une fonction, où I est une partie. Si f est strictement monotone, alors tout nombre réel a au plus un antécédent par f.*

Démonstration. Supposons par exemple que f est strictement croissante sur I. Soit b un nombre réel. Faisons l'hypothèse que b a un antécédent $a \in I$. Soit x un élément de I tel que $x < a$; puisque f est strictement croissante, on a $f(x) < f(a) = b$, donc $f(x)$ n'est pas égal à b, donc x n'est pas un antécédent de b. De même, si $x > a$, alors $f(x) > f(a) = b$ et x n'est pas un antécédent de b. Le seul antécédent de b est donc a. Tout nombre réel a donc zéro ou un antécédent par f. ■

Partition définie par une application

Soit $f : E \to F$ une application. Pour tout élément $b \in F$, notons $A(b)$ l'ensemble des antécédents de b par f : on a donc

$$A(b) = \{x \in E \mid f(x) = b\}.$$

Voici des propriétés générales de ces parties $A(b)$ de E.

a) *La réunion de toutes les parties $A(b)$ est l'ensemble E.*

En effet, si a est un élément quelconque de E, alors a est un antécédent de l'élément $b = f(a)$, donc a appartient à $A(b)$.

b) *Si b et b' sont deux éléments différents appartenant à F, alors l'intersection $A(b) \cap A(b')$ ne contient aucun élément.*

Supposons en effet que b et b' sont des éléments de F et qu'il existe au moins un élément $x \in A(b) \cap A(b')$; puisque $x \in A(b)$, on a $f(x) = b$ et puisque $x \in A(b')$, on a de même $f(x) = b'$; on en déduit $b = b'$. Par conséquent, si $b \neq b'$, il n'y a aucun élément dans l'intersection $A(b) \cap A(b')$.

Certaines parties $A(b)$ peuvent ne contenir aucun élément ; si l'on supprime ces parties vides, on obtient une *partition de l'ensemble E*, c'est-à-dire par définition, des parties non vides, deux à deux disjointes et dont la réunion est l'ensemble E.

Définition

Supposons que f est une application à valeurs réelles définie sur une partie de $\mathbb{R}^2$ ou de $\mathbb{R}^3$. Pour tout nombre réel k, l'ensemble $A(k)$ s'appelle la *ligne* ou la *surface de niveau k* de l'application f.

Sur la carte topographique d'une région montagneuse, les lignes de niveau joignent les points qui sont à la même altitude : par exemple, à proximité d'un col, les lignes de niveau ont l'allure typique représentée figure 4 ; les figures 1, 2 et 3 montrent les coupes de terrain correspondantes. Sur une carte marine, les lignes de sonde sont les lignes de niveau pour la fonction profondeur.

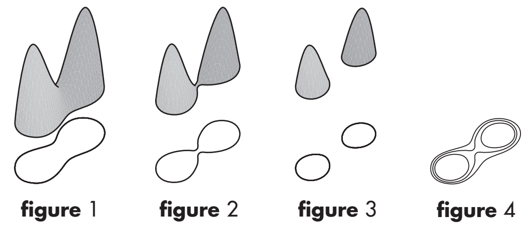
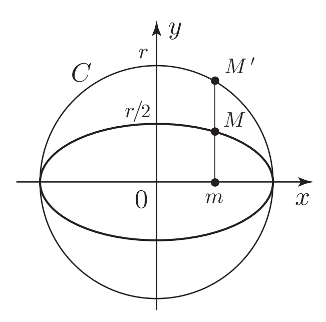
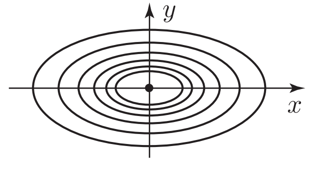


figure 1 **figure** 2 **figure** 3 **figure** 4

Exemple 1. Pour tout $(x, y) \in \mathbb{R}^2$, posons $f(x, y) = x^2 + 4y^2$. Pour cette fonction, la ligne de niveau k a pour équation $x^2 + 4y^2 = k$. Si $k < 0$, $A(k)$ est l'ensemble vide. L'ensemble $A(0)$ est réduit au point $(0, 0)$. Supposons $k = r^2$, où $r > 0$, et notons C le cercle de centre $(0,0)$ et de rayon r. On a les équivalences $(x,y) \in A(k) \Longleftrightarrow x^2 + (2y)^2 = r^2 \Longleftrightarrow (x, 2y) \in C$.
Notons M le point de coordonnées (x,y), M' le point de coordonnées $(x, 2y)$ et m le point de coordonnées $(x, 0)$. Comme M est le milieu de mM', on en déduit une construction de la courbe $A(k)$: faire parcourir à M' le cercle C et prendre les milieux des segments mM', où m est le projeté de M' sur l'axe des abscisses. La courbe $A(k)$ est une ellipse de centre O et d'axes Ox, Oy ; on dit que C est son cercle directeur. Pour des valeurs différentes de k, ces ellipses sont disjointes et quand k parcourt $[0, +\infty[$, la réunion des $A(k)$ est le plan tout entier.

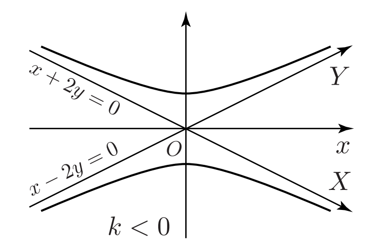


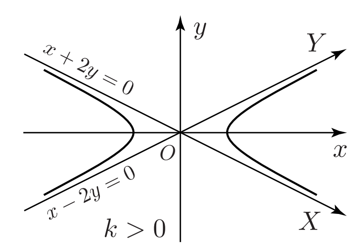
construction de l'ellipse
$x^2 + 4y^2 = r^2$



des lignes de niveau de f

Exemple 2. La fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x, y) = x^2 - 4y^2$ a pour ligne de niveau k la courbe d'équation $x^2 - 4y^2 = k$. Si $k = 0$, on obtient la réunion des deux droites d'équation $x + 2y = 0$ et $x - 2y = 0$, car $x^2 - 4y^2 = (x + 2y)(x - 2y)$.
Supposons $k \neq 0$ et choisissons un repère centré à l'origine O dont les axes OX et OY sont portés par les droites d'équation $x + 2y = 0$ et $x - 2y = 0$. L'équation de la ligne de niveau dans ce repère devient $XY = c$, où c est une constante qui dépend de k et de l'unité de longueur sur les axes : la courbe est donc une hyperbole d'asymptotes les axes OX et OY ; les branches sont situées dans des quarts de plans opposés qui dépendent du signe de k.





Application bijective

Définition

On dit qu'une application $f : E \to F$ est *bijective*, ou que f est une *bijection*, si tout élément de F possède exactement un antécédent. Pour tout $y \in F$, l'unique antécédent de y se note $f^{-1}(y)$.

Les deux propriétés caractéristiques d'une bijection $f : E \to F$ sont donc :

i) si x et x' sont des éléments de E tels que $f(x) = f(x')$, alors $x = x'$;

ii) tout élément de F s'écrit $f(x)$ pour un certain élément $x \in E$.

Supposons que l'application $f : E \to F$ est bijective. À tout élément $y \in F$, associons l'unique antécédent de y : on définit ainsi une application de F vers E notée $f^{-1} : F \to E$.

Définition

Si $f : E \to F$ est une bijection, l'application de F vers E qui à tout $y \in F$ associe $f^{-1}(y)$ s'appelle la *bijection réciproque* de f et se note $f^{-1} : F \to E$.

Si f est une bijection, alors par définition,

$$\text{pour tout } y \in F, \text{ on a } f\big(f^{-1}(y)\big) = y\,,$$

et puisque tout élément $x \in E$ est antécédent de $f(x)$, on a

$$\text{pour tout } x \in E,\ f^{-1}\big(f(x)\big) = x.$$

On en déduit que si f est une bijection, alors l'application f^{-1} est bijective et la bijection réciproque de f^{-1} est f.

Définition

Si E est un ensemble, l'application *identité de* E est la transformation de E qui à tout $x \in E$ associe x lui-même. On note id_E l'application identité de E.

Par définition, si l'on compose une bijection f et sa bijection réciproque, on obtient l'identité ; précisément, si $f : E \to F$ est une bijection, alors on a

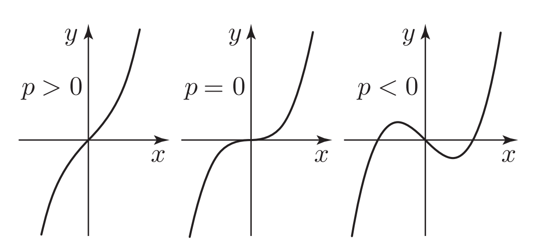
$$f^{-1} \circ f = \mathrm{id}_E :\ E \xrightarrow{\ f\ } F \xrightarrow{\ f^{-1}\ } E \quad \text{et} \quad f \circ f^{-1} = \mathrm{id}_F :\ F \xrightarrow{\ f^{-1}\ } E \xrightarrow{\ f\ } F$$

Exemples

1) La fonction sinus définit une bijection $[-\pi/2, \pi/2] \to [-1, 1]$ (exemples (2) et (3) page 20). De même, pour tout nombre $y \in [-1, 1]$, il existe un unique nombre $x \in [0, \pi]$ tel que $\cos x = y$. La fonction cosinus définit donc une bijection $[0, \pi] \to [-1, 1]$.

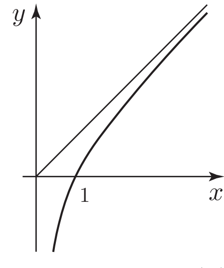
2) La fonction exponentielle prend des valeurs strictement positives et définit une bijection $\exp : \mathbb{R} \to]0, +\infty[$. La bijection réciproque est par définition la fonction logarithme $\ln :]0, +\infty[\to \mathbb{R}$, de sorte que l'on a $\ln = \exp^{-1}$ et $\exp = \ln^{-1}$.

3) Soient p un nombre réel et la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^3 + px$. La dérivée est $3x^2 + p$, donc f est strictement croissante si $p \geqslant 0$; de plus, $f(x)$ tend vers $+\infty$ quand x tend vers $+\infty$, vers $-\infty$ quand x tend vers $-\infty$ et f est continue. On en



déduit que f est bijective si $p \geqslant 0$.
Comme on ne sait pas résoudre en x l'équation $x^3+px=y$, il n'y a pas de formule générale pour exprimer $f^{-1}(y)$.
Dans le cas $p<0$, la fonction est décroissante sur $[\sqrt{-p/3}, \sqrt{-p/3}]$ et croissante ailleurs, donc f n'est pas bijective.
La pente de la tangente à l'origine est toujours p.

4) Montrons que la fonction $h :]0,+\infty[\to \mathbb{R}$ définie par $h(x) = x - \dfrac{1}{x}$ est bijective. Nous devons vérifier que pour tout nombre réel y, l'équation $x - \dfrac{1}{x} = y$ a exactement une solution x dans l'intervalle $]0,+\infty[$. L'équation s'écrit $x^2-1=yx$ ou encore $x^2-yx-1=0$, et les deux racines sont $y \pm \sqrt{y^2+4}$. Il y a une seule racine appartenant à l'ensemble de départ $]0,+\infty[$, c'est $y+\sqrt{y^2+4}$. Donc h est bien une bijection et la bijection réciproque est définie par la formule $h^{-1}(y) = y+\sqrt{y^2+4}$, pour tout $y \in \mathbb{R}$.



graphe de $x \mapsto h(x)$

3.4 Changement de référentiel

Soit $f : E \to E$ une transformation d'un ensemble E. Supposons qu'on dispose d'un ensemble E' et d'une bijection $u : E \to E'$.
Pour tout élément $x \in E$, posons $x' = u(x)$. On a donc $x = u^{-1}(x')$.
Si x et y sont des éléments de E, la relation $y = f(x)$ définit entre les éléments $x' = u(x)$ et $y' = u(y)$ une relation de la forme $y' = f'(x')$: précisément, nous avons

$$y' = u(y) = u\big(f(x)\big) = u \circ f(x) = u \circ f\big(u^{-1}(x')\big) \ , \quad \text{pour tout } x' \in E'.$$

En posant $f' = u \circ f \circ u^{-1} : E' \to E'$, on a ainsi l'équivalence

$$y = f(x) \iff y' = f'(x')\,.$$

Définition
Nous dirons que la bijection $u : E \to E'$ est un *changement de référentiel* ou un *changement de coordonnées* et que la transformation $f' = u \circ f \circ u^{-1} : E' \to E'$ est *transportée de f par le changement de référentiel u.*

$$\begin{array}{ccc} E & \overset{f}{-\!\!-\!\!-\!\!-} & E \\ {\scriptstyle u}\Big| & & \Big|{\scriptstyle u} \\ E' & \underset{f'}{-\!\!-\!\!-\!\!-} & E' \end{array} \qquad\qquad \begin{array}{ccc} x & \overset{f}{-\!\!-\!\!-\!\!-} & y \\ {\scriptstyle u}\Big| & & \Big|{\scriptstyle u} \\ x' & \underset{f'}{-\!\!-\!\!-\!\!-} & y' \end{array}$$

Par un changement de référentiel convenable, la transformation f' peut être plus simple à étudier que la transformation f. Voici des propriétés générales d'un changement de référentiel.

Proposition. *Soit $x_0 \in E$. Posons $x'_0 = u(x_0)$.*

i) *Les itérés $x_1 = f(x_0), \ldots, x_k = f(x_{k-1}), \ldots$ de x_0 par f ont pour image par u les itérés $x'_1 = f(x'_0), \ldots, x'_k = f(x'_{k-1}), \ldots$ de x'_0 par f'.*

ii) *L'élément x_0 est point fixe de f si et seulement si x'_0 est point fixe de f' : autrement dit, les points fixes de f' sont les images par u des points fixes de f.*

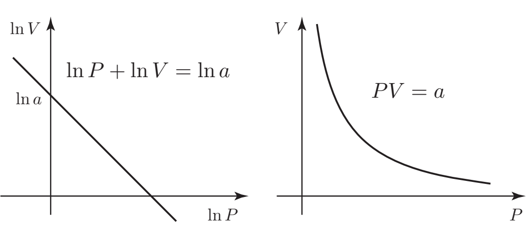
Démonstration. Posons $x'_k = u(x_k)$ pour tout entier $k \geqslant 0$. Raisonnons par récurrence pour montrer que l'on a $x'_k = f'(x'_{k-1})$ pour tout entier $k \geqslant 1$. Puisque $x_1 = f(x_0)$, il vient $x'_1 = u(x_1) = u \circ f(x_0) = u \circ f\big(u^{-1}(x'_0)\big) = f'(x'_0)$. Supposons que k est un entier au moins égal à 1 tel que $x'_k = f'(x'_{k-1})$. Puisque $x_{k+1} = f(x_k)$, on a $x'_{k+1} = u(x_{k+1}) = u \circ f(x_k) = u \circ f\big(u^{-1}(x'_k)\big) = f'(x_k)$, d'où (i). L'élément x_0 est point fixe de f si et seulement si $x_1 = x_0$. Puisque u est une bijection, cette égalité équivaut à $u(x_1) = u(x_0)$, c'est-à-dire à $x'_1 = x'_0$, et cette dernière relation signifie que x'_0 est point fixe de f'. ■

Exemple 1. Posons $\mathbb{R}_+^* =]0, +\infty[$ et considérons la transformation $f : \mathbb{R}_+^* \to \mathbb{R}_+^*$ définie par $f(x) = ax^r$, où a est un nombre positif donné et r un nombre rationnel. Changeons de référentiel au moyen de la fonction logarithme $\ln : \mathbb{R}_+^* \to \mathbb{R}$ qui est bien une bijection. Pour tout $x \in \mathbb{R}_+^*$, posons $y = f(x)$, $x' = \ln x$ et $y' = \ln y$. Il vient

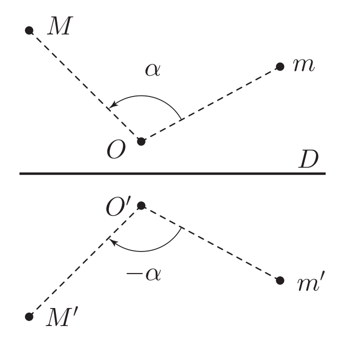
$$y = ax^r \quad \text{et} \quad y' = \ln y = \ln(ax^r) = r \ln x + \ln a = rx' + \ln a$$

d'après les propriétés de la fonction logarithme (page 267). La fonction g qui à x' associe y' est donc simplement la fonction $x' \mapsto rx' + b$, où $b = \ln a$.

Dans la pratique, ce changement de référentiel s'utilise souvent dans les conditions suivantes : une série de mesures portant sur deux variables x et y semble indiquer qu'il y a entre x et y une relation du type $y = ax^r$, mais ni r, ni a ne sont connus. On convertit les mesures x_1, y_1, $x_2, y_2, \ldots$ dans l'échelle logarithmique et l'on marque les points de coordonnées $(\ln x_1, \ln y_1)$, $(\ln x_2, \ln y_2), \ldots$ Supposons que ces points soient à peu près alignés sur une droite ; si r est la pente de la droite et si b est son ordonnée à l'origine, alors $\ln x$ et $\ln y$ sont liés par la relation $\ln y = r \ln x + b$, donc on a $y = e^b x^r$.



Exemple 2. Soient O un point du plan euclidien $\mathscr{P}$ et r la rotation de centre O et d'angle α. Pour tout point m, posons $M = r(m)$.
Soit D une droite et soient m', M', O' les symétriques de m, M, O par rapport à D. Puisque l'angle de la rotation r est α, on a $\widehat{\overrightarrow{Om}, \overrightarrow{OM}} = \alpha$ et comme les angles $\widehat{\overrightarrow{Om}, \overrightarrow{OM}}$ et $\widehat{\overrightarrow{O'm'}, \overrightarrow{O'M'}}$ sont opposés, il vient $\widehat{\overrightarrow{O'm'}, \overrightarrow{O'M'}} = -\alpha$. Le point M' se déduit ainsi de m' par la rotation r' de centre O' et d'angle $-\alpha$.



Notons s la symétrie orthogonale par rapport à D. Par le changement de référentiel s, la rotation r est transportée en la rotation r'.
On a $m=s(m')$ et $M=r(m)$, donc il vient $M'=s(M)=s\circ r(m)=s\circ r\circ s(m')$. Puisque l'application $m'\mapsto M'$ est la rotation r', on en déduit l'égalité de fonctions $s\circ r\circ s=r'$.
Dans le cas particulier où $O\in D$, on a $O=O'$ et donc $r'=r^{-1}$.

Exemple 3. Donnons-nous des nombres réels a, b et p et définissons une transformation $f:\mathbb{R}^2\to\mathbb{R}^2$ en posant

(1) $$f(x,y)=(x',y')\ ,\quad \text{où}\quad x'=ax\quad\text{et}\quad y'=by+px^2.$$

Pour tout $(x,y)\in\mathbb{R}^2$, posons $u(x,y)=(X,Y)$, où

(2) $$X=x\qquad\text{et}\qquad Y=y+sx^2$$

s étant un nombre réel que nous choisirons plus loin. Cela définit une application $(x,y)\mapsto u(x,y)$ de $\mathbb{R}^2$ dans $\mathbb{R}^2$. Puisqu'on a

(3) $$x=X\qquad\text{et}\qquad y=Y-sx^2=Y-sX^2,$$

l'application u est bijective. Pour transformer f par le changement de référentiel u, nous devons exprimer $(X',Y')=u(x',y')$ au moyen de (X,Y). D'après la définition de u donnée en (2), on a

$$X'=x'\qquad\text{et}\qquad Y'=y'+sx'^2.$$

Exprimons x' et y' au moyen de x et y en utilisant (1) et (3) :

$$X'=x'=ax=aX\ ,\quad\text{car } X=x,$$

$$Y'=by+px^2+sx'^2=b(Y-sX^2)+pX^2+s(aX)^2=bY+\big[p-s(b-a^2)\big]X^2$$

Rappelons que nous avons le choix du nombre s : si l'on peut prendre s de manière à annuler le terme en X^2 dans l'expression de Y', alors la relation entre (X,Y) et (X',Y') sera très simple.
L'équation $p-s(b-a^2)=0$ a pour solution $s=\dfrac{p}{b-a^2}$ pourvu que $b\neq a^2$, condition qui dépend de la fonction f.
Supposons $b\neq a^2$. Alors en choisissant pour s la valeur calculée ci-dessus, on obtient les relations linéaires

(4) $$X'=aX\quad\text{et}\quad Y'=bY.$$

qui expriment la transformation f dans les « nouvelles coordonnées » X et Y. Par le changement de référentiel u, f est transportée en $f':(X,Y)\mapsto(aX,bY)$ dont l'expression est plus simple que celle de f. On a l'égalité de fonctions $f'=u\circ f\circ u^{-1}$.
Il est significatif que les nombres a et b, coefficients de x et de y dans les expressions (1), se retrouvent comme coefficients de X et de Y dans (4) : l'explication sera donnée page 389.

3.5 Groupes de transformations

Rappelons qu'une transformation d'un ensemble E est une application de E dans E. Les transformations bijectives sont particulièrement utiles.

Exemples

- Dans l'espace euclidien, une symétrie par rapport à un plan ou une rotation autour d'une droite sont des transformations bijectives.
- Soit F l'ensemble des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ et soit a un nombre réel. Pour toute fonction $f : \mathbb{R} \to \mathbb{R}$, définissons la fonction $f_a : \mathbb{R} \to \mathbb{R}$ en posant $f_a(x) = f(x+a)$ pour tout $x \in \mathbb{R}$. L'application $f \mapsto f_a$ est une transformation de F. Remarquons que si $g = f_a$, alors $g(x-a) = f(x)$ pour tout x, donc $g_{(-a)} = f$: la transformation $f \mapsto f_a$ est donc bijective, la bijection réciproque étant $g \mapsto g_{(-a)}$.

Mettons en évidence les propriétés générales des transformations bijectives d'un ensemble E. Rappelons que l'on note id_E l'application identité de E : c'est une transformation bijective de E.

A) Pour toute transformation f de E, on a $f \circ \mathrm{id}_E = \mathrm{id}_E \circ f = f$.

B) Si f est une transformation bijective de E, la transformation réciproque f^{-1} permet « d'inverser » f : si x et y sont des éléments de E, on a l'équivalence $y = f(x) \Longleftrightarrow x = f^{-1}(y)$. Pour tout $x \in E$, on a $f\big(f^{-1}(x)\big) = x$ et $f^{-1}\big(f(x)\big) = x$, ce qui se traduit par les égalités entre transformations :

$$f \circ f^{-1} = f^{-1} \circ f = \mathrm{id}_E \ .$$

C) On peut toujours composer deux transformations de E. Si f et g sont des transformations bijectives de E, leur composée $g \circ f$ est une transformation bijective de E et la transformation réciproque de $g \circ f$ est $f^{-1} \circ g^{-1}$: pour tout $x \in E$, on a en effet les égalités

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \mathrm{id}_E \circ f = f^{-1} \circ f = \mathrm{id}_E \ .$$

$$E \underset{f^{-1}}{\overset{f}{\rightleftarrows}} E \underset{g^{-1}}{\overset{g}{\rightleftarrows}} E$$

Proposition. *Soient f, g, h des transformations bijectives d'un ensemble E.*

i) *Pour tout entier $n \geqslant 1$, la transformation réciproque de f^n est $(f^n)^{-1} = (f^{-1})^n$.*

ii) *On a les équivalences :*

a) $f = g \Longleftrightarrow f \circ h = g \circ h \Longleftrightarrow h \circ f = h \circ g$.

b) $f = g \Longleftrightarrow f \circ g^{-1} = \mathrm{id}_E \Longleftrightarrow f^{-1} \circ g = \mathrm{id}_E$.

Démonstration. La propriété (i) est vraie si $n = 1$. On a aussi $f^2 = f \circ f$ et donc $(f^2)^{-1} = f^{-1} \circ f^{-1} = (f^{-1})^2$ d'après ce qui précède. La formule générale se démontre en raisonnant par récurrence.
Si $f = g$, alors en composant à droite par la transformation h, on obtient $f \circ h = g \circ h$. Réciproquement, supposons $f \circ h = g \circ h$. En composant à droite par la transformation h^{-1}, nous

obtenons $f \circ h \circ h^{-1} = g \circ h \circ h^{-1}$; puisque $h \circ h^{-1} = \mathrm{id}_E$, il vient $f \circ \mathrm{id}_E = g \circ \mathrm{id}_E$, c'est-à-dire $f = g$. On montre de même l'équivalence $f = g \Longleftrightarrow h \circ f = h \circ g$, d'où (a).
En choisissant $h = g^{-1}$, on obtient $f = g \Longleftrightarrow f \circ g^{-1} = g \circ g^{-1} \Longleftrightarrow f \circ g^{-1} = \mathrm{id}_E$. De même, avec $h = f^{-1}$, on a $f = g \Longleftrightarrow f^{-1} \circ f = f^{-1} \circ g \Longleftrightarrow \mathrm{id}_E = f^{-1} \circ g$. ■

Notation. Soit f une transformation bijective d'un ensemble E. Si n est un entier positif, la transformation $(f^n)^{-1} = (f^{-1})^n$ se note f^{-n}. On a défini ainsi f^p pour tout entier p positif ou négatif. De plus, on pose $f^0 = \mathrm{id}_E$.
Avec cette notation, on vérifie facilement la règle de calcul :

$$f^p \circ f^q = f^{p+q} \quad \text{quels que soient les entiers } p \in \mathbb{Z} \text{ et } q \in \mathbb{Z}.$$

Définition

Soit E un ensemble et soit T un ensemble de transformations bijectives de E. On dit que T est *un groupe de transformations de* E si l'on a les trois propriétés suivantes :

i) la transformation id_E appartient à T ;

ii) si f et g appartiennent à T, la composée $g \circ f$ appartient à T ;

iii) si f appartient à T, la transformation f^{-1} appartient à T. La transformation f^{-1} s'appelle aussi *l'inverse de* f.

Soient T un groupe de transformations et $f \in T$. Pour tout entier $n \geqslant 1$, f^n appartient à T, donc la transformation $(f^n)^{-1} = f^{-n}$ appartient aussi à T. Puisque f^0 est la transformation identité, on a $f^0 \in T$. On en déduit que f^p appartient à T quel que soit l'entier $p \in \mathbb{Z}$.
L'ensemble de toutes les transformations bijectives de E est un groupe de transformations. Voici d'autres exemples.

Exemple 1. Soit O un point du plan euclidien $\mathscr{P}$. L'ensemble des rotations de centre O est un groupe de transformations du plan.

En effet, l'identité de $\mathscr{P}$ est la rotation d'angle nul, la composée de deux rotations de centre O est une rotation de centre O et si r est la rotation de centre O et d'angle α, la transformation réciproque de r est la rotation de centre O et d'angle $-\alpha$.

Exemple 2. Soit f une transformation bijective d'un ensemble E. L'ensemble des transformations f^n, où n parcourt $\mathbb{Z}$, est un groupe de transformations ; on dit que c'est le groupe *engendré par* f.

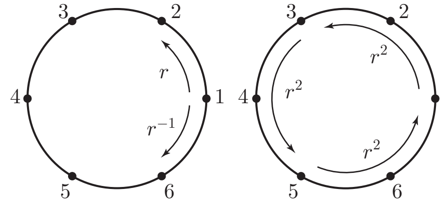
Exemple 3. Soient O un point du plan euclidien $\mathscr{P}$ et r la rotation de centre O et d'angle $\pi/3$. Soit T le groupe engendré par r, c'est-à-dire que T est l'ensemble des rotations r^n, où $n \in \mathbb{Z}$. Le tableau ci-contre indique l'angle des rotations r, r^2, r^3, r^4 et r^5 :

r	r^2	r^3	r^4	r^5
$\pi/3$	$2\pi/3$	π	$4\pi/3$	$5\pi/3$

Puisqu'on a $6(\pi/3)=2\pi$, la rotation r^6 est d'angle nul, donc $r^6=\mathrm{id}_{\mathscr{P}}$. On en déduit $r^7=r\circ r^6=r$, $r^8=r^2$ et aussi $r^5=r^6\circ r^{-1}=r^{-1}$. Plus généralement, si n est un entier de signe quelconque, on a $r^{n+6}=r^n\circ r^6=r^n$. Le groupe engendré par r ne contient donc que les six rotations r^k, où $0\leqslant k\leqslant 5$, de sorte qu'on a $T=\{\mathrm{id}_{\mathscr{P}},r,r^2,r^3,r^4,r^5\}$. Voici le tableau des inverses des éléments de T :

f	$\mathrm{id}_{\mathscr{P}}$	r	r^2	r^3	r^4	r^5
f^{-1}	$\mathrm{id}_{\mathscr{P}}$	r^5	r^4	r^3	r^2	r

- On a $(r^2)^3=r^6=\mathrm{id}_{\mathscr{P}}$, donc $(r^2)^{n+3}=r^{2n+6}=r^{2n}=(r^2)^n$ pour tout entier $n\in\mathbb{Z}$; le groupe engendré par r^2 est donc seulement formé des trois rotations $(r^2)^0=\mathrm{id}_{\mathscr{P}}$, r^2 et $(r^2)^2=r^4$; les transformations r^2 et r^4 sont inverses l'une de l'autre.
- Puisque $(r^3)^2$ est l'identité, on en déduit de même que le groupe engendré par r^3 est l'ensemble à deux éléments $\{\mathrm{id}_{\mathscr{P}},r^3\}$. La transformation r^3 est son propre inverse ; d'ailleurs, l'angle de la rotation r^3 étant égal à π, r^3 est la symétrie par rapport au centre O.



La figure de gauche montre les itérés par r du point 1, numérotés de 2 à 6 : la rotation r les déplace circulairement.

A	1	2	3	4	5	6
$r(A)$	2	3	4	5	6	1

Par la rotation r^2, les points 1, 3, 5 sont déplacés circulairement et les points 2, 4, 6 aussi.

A	1	3	5	2	4	6
$r^2(A)$	3	5	1	4	6	2

A	1	4	2	5	3	6
$r^3(A)$	4	1	5	2	6	3

Exemple 4. Soit a un nombre réel. Pour tout nombre réel t, définissons la transformation f_t de $\mathbb{R}^2$ en posant

$$f_t(x,y)=\big(e^{ta}x,e^{ta}(y+tx)\big)\ ,\quad \text{pour tout } (x,y)\in\mathbb{R}^2.$$

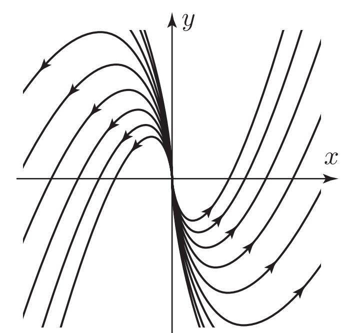
On a $f_0(x,y)=(x,y)$ pour tout (x,y), donc f_0 est la transformation identité de $\mathbb{R}^2$. Posons $(X,Y)=f_t(x,y)$, c'est-à-dire $X=e^{ta}x$ et $Y=e^{ta}(y+tx)$.
Si t' est un nombre réel, il vient $Y+t'X=e^{ta}(y+tx)+t'e^{ta}x=e^{ta}\big(y+(t'+t)x\big)$ et

$$\begin{aligned}(f_{t'}\circ f_t)(x,y)=f_{t'}(X,Y)&=\big(e^{t'a}X\,,\,e^{t'a}(Y+t'X)\big)\\&=\Big(e^{t'a}e^{ta}x\,,\,e^{t'a}e^{ta}\big(y+(t'+t)x\big)\Big)\\&=\Big(e^{(t'+t)a}x\,,\,e^{(t'+t)a}\big(y+(t'+t)x\big)\Big)=f_{t'+t}(x,y)\,.\end{aligned}$$

On a donc l'égalité $f_{t'}\circ f_t=f_{t'+t}$. Cela montre que la composée de deux transformations du type f_t est du même type. En choisissant $t'=-t$, il vient l'égalité $f_{-t}\circ f_t=f_0=\mathrm{id}_{\mathbb{R}^2}$: ainsi chaque transformation f_t est une bijection et l'on a $(f_t)^{-1}=f_{-t}$.

L'ensemble des transformations f_t, où t parcourt $\mathbb{R}$, est donc un groupe de transformations de $\mathbb{R}^2$. Remarquons que l'on a $f_{t'} \circ f_t = f_t \circ f_{t'}$, car $t' + t = t + t'$: on peut donc composer les transformations f_t dans l'ordre qu'on veut.
Puisqu'on a toujours $f_t(0,0) = (0,0)$, le point $(0,0)$ est laissé fixe par toutes les transformations f_t.
Soit (p,q) un point de $\mathbb{R}^2$ différent de $(0,0)$. Quand t varie, les points $(x_t, y_t) = f_t(p,q) = \big(pe^{ta}, (q+pt)e^{ta}\big)$ décrivent une courbe passant par (p,q), car $(p,q) = f_0(p,q)$.
Supposons par exemple $a > 0$. Quand t tend vers $-\infty$, les deux coordonnées x_t et y_t tendent vers 0, car $\lim\limits_{t\to-\infty} e^{ta} = \lim\limits_{t\to-\infty} te^{ta} = 0$: cela veut dire que le point (x_t, y_t) tend vers l'origine quand t tend vers $-\infty$. Quand t tend vers $+\infty$, les valeurs absolues des coordonnées x_t et y_t tendent vers l'infini.
Sur la figure ci-contre, nous avons représenté quelques-unes de ces courbes ; la flèche indique le sens de parcours quand t augmente. Si l'on suppose que t est le temps, le point $M = (x_t, y_t)$ de la courbe a pour vitesse à l'instant t le vecteur $\frac{dM}{dt} = \left(\frac{dx}{dt}, \frac{dy}{dt}\right) = \left(\frac{d}{dt}[pe^{ta}], \frac{d}{dt}\left[(q+pt)e^{ta}\right]\right)$. Il vient donc

$$\left(\frac{dx}{dt}, \frac{dy}{dt}\right) = \left(ape^{ta}, ae^{ta}(q+pt) + pe^{ta}\right) = (ax, ay + x)$$



Ainsi le mouvement de M est régi par le système d'équations différentielles $\begin{cases} \dot{x} = ax \\ \dot{y} = ay + x \end{cases}$, où $\dot{x}$ et $\dot{y}$ désignent comme d'habitude la dérivée par rapport au temps. Remarquons que le vecteur vitesse du point M ne dépend que des coordonnées de M. Les équations différentielles de ce type seront traitées au chapitre 16.

Exercices

@ **1. Un exemple de suite périodique.** Pour tout entier $n \geqslant 1$, notons r_n le reste de la division euclidienne de 2^n par 15.

a) Calculer r_1, r_2, r_3 et r_4, puis montrer que l'on a $r_{n+4} = r_n$ pour tout n.

b) En déduire la valeur de r_n, quel que soit l'entier $n \geqslant 1$.

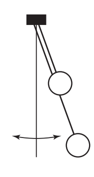
@ **2. Meilleure approximation de $\sqrt{2}$ par des fractions**

a) Montrer que la partie entière de $\sqrt{2}$ est égale à 1.

b) Montrer que l'on a $\sqrt{2} = 1 + \dfrac{1}{1+\sqrt{2}}$ et $1 + \sqrt{2} = 2 + \dfrac{1}{1+\sqrt{2}}$.

c) En déduire les quatre premières fractions de la meilleure approximation de $\sqrt{2}$.

d) On dispose de deux pendules simples, constitués chacun d'une masse suspendue à l'extrémité d'un fil non pesant ; pour de petites oscillations, la période est proportionnelle à la racine carrée de la longueur du fil. La longueur de l'un des pendules est double de l'autre. En position de départ, les pendules sont écartés de leur position d'équilibre d'un même petit angle, puis on les lâche au même instant.



(i) Montrer que les pendules ne reviendront jamais exactement dans la configuration de départ.

(ii) On observe qu'après sept oscillations complètes du petit pendule, les deux pendules sont presque dans la position de départ. Expliquer pourquoi.

(ii) Ces retours simultanés presque à la position de départ se reproduisent avec une précision croissante après $n_1, n_2, n_3, \ldots$ oscillations complètes du petit pendule. Calculer les entiers n_1, n_2 et n_3.

@ **3. Itération affine.** Soient a et b des nombres réels et soit $f : \mathbb{R} \to \mathbb{R}$ la fonction définie par $f(x) = ax + b$. Donnons-nous un nombre x_0 et notons x_n les itérés de x_0 par f : ils sont définis par la relation $x_{n+1} = f(x_n)$, pour tout entier $n \geqslant 0$.

a) Montrer que si $|a| < 1$, la suite (x_n) a pour limite le point fixe ω de f.

b) On suppose $x_0 \neq \omega$ et $|a| > 1$. Montrer que le rapport x_n/a^n tend vers $x_0 - \omega$ quand n tend vers l'infini. En déduire que $|x_n|$ tend vers $+\infty$.

4. On place un capital c à un taux d'intérêt i ; le montant de la prime versée annuellement est b. Posons $r = 1 + i$ et notons c_n la somme disponible après n années. Montrer que l'on a $c_{n+1} = rc_n + b$. En déduire que $c_n = cr^n + (b/i)(r^n - 1)$.

@ **5. Un modèle d'offre et de demande.** Dans certains secteurs économiques (comme l'agriculture), le prix p des biens pendant une période est fonction de la quantité q de biens consommés et la production Q pendant cette période est fonction du prix p' pratiqué pendant la période précédente. Supposons qu'en utilisant des unités convenables, ces fonctions sont assez bien représentées par les formules $p = 80 - (1/2)q$ et $Q = (1/3)p' + 20$ (remarquer que p est fonction décroissante de q et que Q est fonction croissante de p'). Supposons aussi que l'équilibre $q = Q$ est réalisé. Notons p_n et q_n le prix et la consommation pendant la n-ième période.

a) Montrer qu'on a la relation $p_{n+1} = 70 - (1/6)p_n$. En déduire l'égalité $p_n = (20 - q_0/2)(-1/6)^n + 60$. Vers quelle limite tendent les prix ?

b) On suppose $q_0 = 20$. Pour visualiser l'évolution des prix, dessiner, dans un repère du plan, les points de coordonnées (n, p_n), pour $0 \leqslant n \leqslant 6$. Observer que la fluctuation s'amortit.

6. Soient p un nombre réel et $f : \mathbb{R} \to \mathbb{R}$ la fonction définie par $f(x) = x^3 - px$, où $p > 0$.

a) Étudier les variations de f et dessiner le graphe.

b) Si b est un nombre réel, déterminer le nombre d'antécédents de b par f (discuter selon les valeurs b et de p).

@ **7. a)** Montrer que les fonctions $t \mapsto t - \sin t$ et $t \mapsto t - \cos t$ sont des bijections strictement croissantes de $\mathbb{R}$ dans $\mathbb{R}$ (utiliser la dérivée).

b) Supposons que deux quantités réelles x et y sont reliées par la relation $x + y = \sin x + \cos y$. Montrer que y est fonction strictement décroissante de x. En déduire que l'équation $2x = \sin x + \cos x$ a au plus une solution.

c) Montrer que l'équation $2x = \sin x + \cos x$ a exactement une solution, comprise entre $\pi/6$ et $\pi/4$ (étudier la fonction $2x - \sin x - \cos x$ sur l'ensemble $\mathbb{R}$: elle est croissante).

@ **8. Des lignes de niveau sonore.** Sur un terrain plan, on pose des amplificateurs a_1 et a_2 en des points A_1 et A_2 distants de 20m. Pour chacun de ces amplificateurs, l'intensité sonore reçue en un point est inversement proportionnelle au carré de la distance à l'amplificateur.
Quelle est la courbe formée des points où l'intensité reçue de a_1 est k fois l'intensité reçue de a_2, k étant un nombre positif donné ? Comment évolue cette courbe lorsque k devient grand ?

Choisir un repère orthonormé d'origine le milieu de (A_1, A_2), avec l'axe des x porté par la droite A_1A_2. Si $k \neq 1$, l'équation trouvée est de la forme $x^2 + y^2 - 2ux + v = 0$ et la courbe est un cercle centré en un point de la droite A_1A_2 ; si $k = 1$, la solution est la médiatrice de (A_1, A_2).

9. Pour chacune des fonctions $f : \mathbb{R}^2 \to \mathbb{R}$ suivantes, représenter sur un même dessin les lignes de niveau indiquées :

a) $f(x,y) = x^2y^2$, niveaux 0, 1 et 4 **b)** $f(x,y) = |x|+|y|$, niveaux 0, 1 et 2
c) $f(x,y) = |y-1|+x$, niveaux -1, 0, 1 **d)** $f(x,y) = |y-1|+x^2$, niveaux 0, 1, 2

@ **10. Utilisation d'un changement de référentiel**

Pour tout $x \geqslant 0$, posons $u(x) = \dfrac{x}{1+x}$ et $f(x) = \dfrac{x^2}{1+2x}$.

a) Montrer que la fonction u est une bijection de $[0, +\infty[$ sur $[0, 1[$.

b) Montrer que f est une transformation de l'intervalle $[0, +\infty[$.

c) Soit $g : [0,1[\to [0,1[$ la fonction transportée de f par le changement de référentiel u. Montrer que l'on a $g(x') = x'^2$ pour tout $x' \in [0, 1[$.

d) Quelle est la limite des itérés par g d'un nombre $x'_0 \in [0, 1[$? Quelle est la limite des itérés par f d'un nombre $x_0 \geqslant 0$?

e) Montrer que f est une transformation bijective de l'intervalle $[0, +\infty[$. Quelle est la limite des itérés de x_0 par f^{-1} ?

11. Effet « zoom ». On a fait dessiner, par un ordinateur, le graphe d'une fonction f sur l'intervalle $[-a, a]$ (où $a > 0$). Cette fonction est telle que $f(0) = 0$. On veut faire subir au dessin un effet « zoom » d'un facteur $k > 1$ en centrant l'opération sur l'origine. Cela revient à changer de référentiel au moyen de la bijection $x \mapsto kx$. Quelle fonction doit-on demander à l'ordinateur de dessiner pour obtenir le résultat ?

12. Reprenons l'exemple **4** page 29 dans le cas $a = 1$. Étant donné un point (x_0, y_0), nous avons dessiné (pour $a = 1$) la courbe C décrite par les points de coordonnées $x_t = e^t x_0$, $y_t = e^t(y_0 + tx_0)$ quand t varie. On se propose, dans cet exercice, de justifier l'allure de la courbe. Pour simplifier, supposons $x_0 = 1$.

a) Montrer qu'un point (x, y) est sur la courbe C si et seulement si l'on a $x > 0$ et $y = y_0 x + x \ln x$.

b) Étudier la fonction $x \mapsto y_0 x + x \ln x$. Représenter sur un même dessin le graphe de cette fonction pour les valeurs $y_0 = 0$, $y_0 = 1$, $y_0 = -1$.

@13. Un groupe de transformations géométriques. Soit $\mathscr{P}$ le plan euclidien muni d'un repère orthonormé $(O; \vec{i}, \vec{j})$. On note t la translation de vecteur $\vec{i}$ et s la symétrie par rapport à l'axe des ordonnées. Si $a \in \mathbb{R}$, notons D_a la droite parallèle à l'axe des ordonnées et passant par le point $(a, 0)$.

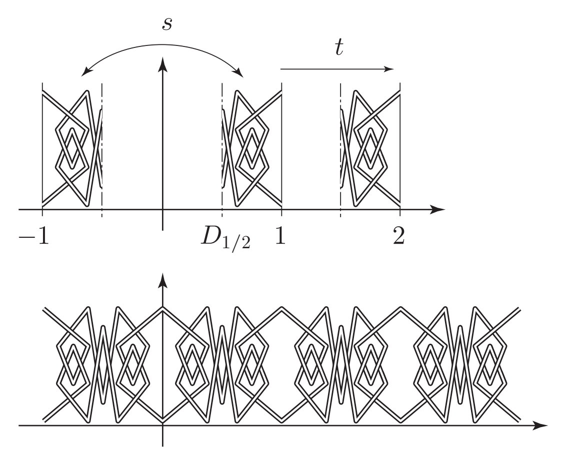
a) Pour tout point M de coordonnées (x, y), calculer les coordonnées des points $t(M)$ et $s(M)$.

b) Montrer que $s \circ s$ est l'identité, que $t \circ s$ est la symétrie par rapport à la droite $D_{1/2}$ et que $t^2 \circ s$ est la symétrie par rapport à la droite D_1.

c) Montrer que l'on a $s \circ t = t^{-1} \circ s$.

d) En déduire les relations $s \circ t^p = t^{-p} \circ s$ pour tout entier $p \in \mathbb{Z}$ et $(s \circ t^p) \circ (s \circ t^q) = t^{q-p}$ pour tous entiers p et q.

e) En déduire que l'ensemble $T = \{t^p \mid p \in \mathbb{Z}\} \cup \{s \circ t^p \mid p \in \mathbb{Z}\}$ est un groupe de transformations du plan. Vérifier que T contient les symétries par rapport aux droites D_p et $D_{p/2}$, quel que soit l'entier $p \in \mathbb{Z}$.



le groupe T génère une frise

Chapitre 2

Nombres complexes et polynômes

1. Les nombres complexes

Ajoutons à l'ensemble des nombres réels un symbole $\boldsymbol{i}$ et étendons à cet ensemble les opérations somme et produit : on forme ainsi les expressions $r_0+r_1\boldsymbol{i}+r_2\boldsymbol{i}^2+\cdots+r_p\boldsymbol{i}^p$, où les coefficients r_k sont des nombres réels. La somme et le produit de deux telles expressions est de la même forme.

Convenons que le symbole $\boldsymbol{i}$ *vérifie la relation* $\boldsymbol{i}^2=-1$.

L'ensemble des expressions $r_0+r_1\boldsymbol{i}+r_2\boldsymbol{i}^2+\cdots+r_p\boldsymbol{i}^p$, muni de la somme et du produit, se note $\mathbb{C}$ et s'appelle *l'ensemble des nombres complexes*. La relation $\boldsymbol{i}^2=-1$ permet de simplifier l'expression d'un nombre complexe : on a $\boldsymbol{i}^3=(\boldsymbol{i}^2)\,\boldsymbol{i}=-\boldsymbol{i}$ et $\boldsymbol{i}^4=(-1)^2=1$, donc par exemple $1-2\boldsymbol{i}+3\boldsymbol{i}^3+4\boldsymbol{i}^5=1-2\boldsymbol{i}+3(-\boldsymbol{i})+4\boldsymbol{i}=1-\boldsymbol{i}$. Plus généralement, pour tout entier $n\geqslant 0$, on a $\boldsymbol{i}^{2n}=(\boldsymbol{i}^2)^n=(-1)^n$ et $\boldsymbol{i}^{2n+1}=(\boldsymbol{i}^{2n})\boldsymbol{i}=(-1)^n\boldsymbol{i}$. Il s'ensuit que tout élément de $\mathbb{C}$ s'écrit $a+b\boldsymbol{i}$, avec a et b des nombres réels.

1.1 Règles de calcul sur les nombres complexes

Voici les règles concernant les expressions $a+b\boldsymbol{i}$, où a et b sont des nombres réels.

i) $(a+b\boldsymbol{i})+(a'+b'\boldsymbol{i})=(a+a')+(b+b')\boldsymbol{i}$ *et* $(a+b\boldsymbol{i})(a'+b'\boldsymbol{i})=(aa'-bb')+(ab'+ba')\boldsymbol{i}$.

ii) *On a l'équivalence* $a+b\boldsymbol{i}=0 \iff (a=0$ *et* $b=0)$.

Démontrons le sens direct de (ii). Supposons que a et b sont des nombres réels tels que $a+b\boldsymbol{i}=0$, ou encore $a=-b\boldsymbol{i}$. En élevant au carré, on obtient $a^2=(-b\boldsymbol{i})^2=b^2\boldsymbol{i}^2=-b^2$. Mais puisque a et b sont des nombres réels, a^2 est positif ou nul et $-b^2$ est négatif ou nul, donc $a=b=0$.

Définitions

Soit $z = a + b\,\boldsymbol{i}$ un nombre complexe, où a et b sont des nombres réels.

- Le nombre a, appelé *partie réelle* de z, se note $\mathcal{R}e(z)$; le nombre b, appelé *partie imaginaire* de z, se note $\mathcal{I}m(z)$.
- Le *conjugué* de z est le nombre complexe $\bar{z} = a - b\,\boldsymbol{i}$. Le *module* de z est le nombre réel positif ou nul $|z| = \sqrt{a^2 + b^2}$.

On a les relations : $2\,\mathcal{R}e(z) = z + \bar{z}$, $2\,\boldsymbol{i}\,\mathcal{I}m(z) = z - \bar{z}$, $|z| = |\bar{z}|$, $z\bar{z} = |z|^2$
les équivalences : $(\ |z| = 0 \iff z = 0\ )$, $(\ z \in \mathbb{R} \iff z = \bar{z}\ )$
et les inégalités : $|\mathcal{R}e(z)| \leqslant |z|$ et $|\mathcal{I}m(z)| \leqslant |z|$.

Pour montrer ces inégalités, il suffit de remarquer que si $z = a + b\,\boldsymbol{i}$, alors $a^2 \leqslant a^2 + b^2$ donc $|a| \leqslant \sqrt{a^2 + b^2}$, et de même $|b| \leqslant \sqrt{a^2 + b^2}$.

Si z est un nombre réel, le module de z est simplement la valeur absolue.

Inverse d'un nombre complexe non nul. Si z est un nombre complexe non nul, le module de z est un nombre réel non nul ; en divisant par $|z|^2$ l'égalité $z\bar{z} = |z|^2$, on obtient $z\left[\dfrac{1}{|z|^2}\,\bar{z}\right] = 1$: cela signifie que le nombre complexe $\dfrac{1}{|z|^2}\,\bar{z}$ est l'inverse de z pour la multiplication.

Si $z = a + b\boldsymbol{i}$ est un nombre complexe non nul, l'inverse de z pour la multiplication est le nombre complexe $z^{-1} = \dfrac{1}{z} = \dfrac{\bar{z}}{|z|^2} = \dfrac{a}{a^2 + b^2} - \dfrac{b}{a^2 + b^2}\,\boldsymbol{i}$.

Calculs sur le conjugué et le module. *Pour tous nombres complexes z et z', on a*

i) $\overline{z + z'} = \bar{z} + \overline{z'}$, $\overline{zz'} = \bar{z}\,\overline{z'}$ *et si* $z \neq 0$, $\overline{\left(\dfrac{z'}{z}\right)} = \dfrac{\overline{z'}}{\bar{z}}$.

ii) $|zz'| = |z||z'|$ *et si* $z \neq 0$, $\left|\dfrac{z'}{z}\right| = \dfrac{|z'|}{|z|}$.

iii) $|z + z'| \leqslant |z| + |z'|$ *(inégalité triangulaire)*

Démonstration

i) Si $z = a + b\boldsymbol{i}$ et $z' = a' + b'\,\boldsymbol{i}$, alors $\overline{z + z'} = a + a' - b\boldsymbol{i} - b'\,\boldsymbol{i} = \bar{z} + \overline{z'}$ et $\bar{z}\,\overline{z'} = (a - b\boldsymbol{i})(a' - b'\,\boldsymbol{i}) = (aa' - bb') - (ab' + ba')\,\boldsymbol{i} = \overline{zz'}$. Pour la troisième égalité, il suffit de montrer que pour $z \neq 0$, le conjugué de $\dfrac{1}{z}$ est $\dfrac{1}{\bar{z}}$, ce qui résulte de la formule $\dfrac{1}{a + b\,\boldsymbol{i}} = \dfrac{a}{a^2 + b^2} - \dfrac{b}{a^2 + b^2}\,\boldsymbol{i}$.

ii) Puisque $|z|^2 = z\bar{z}$ et $|z'|^2 = z'\,\overline{z'}$, il vient

$$(|z||z'|)^2 = |z|^2|z'|^2 = z\bar{z}z'\,\overline{z'} = (zz')(\bar{z}\,\overline{z'}) = (zz')(\overline{zz'}) = |zz'|^2\,.$$

Si $z \neq 0$, alors on a $1 = |1| = \left|z\dfrac{1}{z}\right| = |z|\left|\dfrac{1}{z}\right|$, donc $\left|\dfrac{1}{z}\right| = \dfrac{1}{|z|}$.

iii) On a

$$\begin{aligned} |z + z'|^2 &= (z + z')(\bar{z} + \overline{z'}) = z\bar{z} + z\overline{z'} + z'\bar{z} + z'\,\overline{z'} \\ &= |z|^2 + 2\,\mathcal{R}e(\bar{z}z') + |z'|^2\ , \quad \text{car } z\overline{z'} = \overline{\bar{z}z'}. \end{aligned}$$

Par suite,

$$\begin{aligned}|z+z'|^2 &\leqslant |z|^2 + 2\,|\mathscr{R}e(\bar{z}z')| + |z'|^2 \quad \text{car pour tout nombre réel } x\text{, on a } x \leqslant |x| \\ &\leqslant |z|^2 + 2|z||z'| + |z'|^2 \qquad \text{car } |\mathscr{R}e(\bar{z}z')| \leqslant |\bar{z}z'| = |\bar{z}||z'| = |z||z'| \\ &\leqslant (|z|+|z'|)^2\,.\end{aligned}$$

Puisque le module est un nombre réel positif ou nul, on en déduit l'inégalité triangulaire en prenant les racines carrées. ■

Exemple. Supposons qu'un nombre complexe z vérifie l'inégalité $|z-a| < a$, où a est un nombre réel strictement positif. Puisque $\mathscr{R}e(a) = a$, on a alors $|\mathscr{R}e(z) - a| = |\mathscr{R}e(z-a)| \leqslant |z-a| < a$, ou encore $-a < \mathscr{R}e(z) - a < a$: cela implique $\mathscr{R}e(z) > 0$.

Argument d'un nombre complexe. Rappelons la propriété suivante :

si a et b sont des nombres réels tels que $a^2+b^2=1$, il existe un unique nombre réel $t \in [0, 2\pi[$ tel que $a = \cos t$ et $b = \sin t$.

Puisqu'un nombre complexe de module 1 s'écrit $a + b\boldsymbol{i}$, où a et b sont des nombres réels vérifiant $a^2+b^2=1$, on en déduit que tout nombre complexe de module 1 est de la forme $\cos t + \boldsymbol{i}\sin t$.
Si z est un nombre complexe non nul, le nombre $\frac{z}{|z|}$ est de module 1, donc s'écrit $\frac{z}{|z|} = \cos t + \boldsymbol{i}\sin t$, où $t \in [0, 2\pi[$.

Définition

Tout nombre complexe $z \neq 0$ s'écrit de manière unique $z = |z|(\cos t + \boldsymbol{i}\sin t)$, où $t \in [0, 2\pi[$. Le nombre t s'appelle *l'argument* de z et se note $\mathrm{Arg}(z)$.

- Des nombres complexes non nuls sont égaux si et seulement s'ils ont même module et même argument.
- Soit z un nombre complexe non nul. Le nombre z est réel positif si et seulement si $\mathrm{Arg}(z) = 0$. Le nombre z est réel négatif si et seulement si $\mathrm{Arg}(z) = \pi$.

Formules de Moivre. *Pour tous nombres réels t et t', on a*

i) $\big(\cos t + \boldsymbol{i}\sin t\big)\big(\cos t' + \boldsymbol{i}\sin t'\big) = \cos(t+t') + \boldsymbol{i}\sin(t+t')$.

ii) $\big(\cos t + \boldsymbol{i}\sin t\big)^{-1} = \cos t - \boldsymbol{i}\sin t$.

iii) $\big(\cos t + \boldsymbol{i}\sin t\big)^{n} = \cos nt + \boldsymbol{i}\sin nt$, *pour tout entier $n \in \mathbb{Z}$.*

Démonstration. La première égalité est une transcription des formules de trigonométrie $\cos(t+t') = \cos t\cos t' - \sin t\sin t'$ et $\sin(t+t') = \sin t\cos t' + \cos t\sin t'$. Le nombre $\cos t + \boldsymbol{i}\sin t$ étant de module 1, son inverse est égal à son conjugué $\cos t - \boldsymbol{i}\sin t$. La troisième formule se démontre par récurrence pour $n \geqslant 1$ en utilisant (i) ; on en déduit la formule pour $n < 0$ en appliquant (ii) et les relations $\cos(-nt) = \cos(nt)$ et $\sin(-nt) = -\sin(nt)$. ■

Pour pratiquer la dernière formule de Moivre, on utilise le développement de $(a+b)^n$ par la formule du binôme de Newton (page 61).

Argument d'un produit ou d'un quotient. *Si des nombres complexes non nuls z et z' ont pour argument t et t', alors*

- *le produit zz' a pour argument $t + t'$ modulo 2π,*
- *le quotient $\frac{z}{z'}$ a pour argument $t - t'$ modulo 2π.*

1.2 Exponentielle d'un nombre complexe

Pour tout nombre réel x, on sait définir l'exponentielle de x, notée $\exp x$ ou encore e^x. On a $e^0 = 1$, $e^x > 0$ et $e^x e^{x'} = e^{x+x'}$ pour tous nombres réels x, x'.

Définition

Si $z = x + y\boldsymbol{i}$ est un nombre complexe, où x et y sont réels, le nombre complexe $\exp z = (\exp x)(\cos y + \boldsymbol{i} \sin y)$ s'appelle *l'exponentielle de z* ; on le note aussi e^z.

Cette notation est justifiée, car si z est un nombre réel x, alors sa partie imaginaire y est nulle, on a $\cos y + \boldsymbol{i} \sin y = 1$ et $\exp z$ est égal à l'exponentielle réelle de x. D'après la définition, on a pour tous nombres réels x et y, les relations

- $e^{\boldsymbol{i}\,y} = \cos y + \boldsymbol{i} \sin y$, $e^{x+\boldsymbol{i}\,y} = e^x\, e^{\boldsymbol{i}\,y}$
- $\cos y = \mathcal{R}e(e^{\boldsymbol{i}\,y}) = \frac{1}{2}(e^{\boldsymbol{i}\,y} + e^{-\boldsymbol{i}\,y})$, $\sin y = \mathcal{I}m(e^{\boldsymbol{i}\,y}) = \frac{1}{2\boldsymbol{i}}(e^{\boldsymbol{i}\,y} - e^{-\boldsymbol{i}\,y})$
- $e^{\boldsymbol{i}\,\pi/2} = \boldsymbol{i}$, $e^{\boldsymbol{i}\,\pi} = -1$, $e^{2\boldsymbol{i}\,\pi} = 1$.
- $e^{\boldsymbol{i}(y+\pi)} = -e^{\boldsymbol{i}\,y}$, $e^{\boldsymbol{i}(y+2\pi))} = e^{\boldsymbol{i}\,y}$

Exemples. On a $1 + \boldsymbol{i} = \sqrt{2}\, e^{\boldsymbol{i}\,\pi/4}$, $\sqrt{3} + \boldsymbol{i} = 2e^{\boldsymbol{i}\,\pi/6}$ et $1 + \boldsymbol{i}\,\sqrt{3} = 2e^{\boldsymbol{i}\,\pi/3}$.

Propriétés de l'exponentielle

1) *Pour tous nombres complexes z et z', on a $e^z e^{z'} = e^{z+z'}$.*
2) *Pour tout nombre complexe z, on a $\overline{e^z} = e^{\overline{z}}$, $|e^z| = e^{\mathcal{R}e(z)}$ et $\operatorname{Arg}(e^z) = \mathcal{I}m(z)$ modulo 2π.*
3) *Pour tous nombres complexes z et z', on a l'équivalence $e^{z'} = e^z \iff z' = z + 2k\pi\boldsymbol{i}$, où $k \in \mathbb{Z}$.*

Démonstration

1) En posant $z = x + \boldsymbol{i}\,y$ et $z' = x' + \boldsymbol{i}\,y'$, où x, y, x', y' sont réels, il vient

$$\begin{aligned} e^z e^{z'} &= (\exp x)(\cos y + \boldsymbol{i} \sin y)(\exp x')(\cos y' + \boldsymbol{i} \sin y') \\ &= (\exp x)(\exp x')(\cos y + \boldsymbol{i} \sin y)(\cos y' + \boldsymbol{i} \sin y') \\ &= \exp(x + x')\big[\cos(y + y') + \boldsymbol{i} \sin(y + y')\big] = e^{z+z'} \quad , \quad \text{d'après la formule de Moivre.} \end{aligned}$$

2) Posons $z = x + \boldsymbol{i}\,y$, où $x = \mathcal{R}e(z)$ et $y = \mathcal{I}m(z)$. On a l'égalité

$$\overline{e^{\boldsymbol{i}\,y}} = \overline{\cos y + \boldsymbol{i} \sin y} = \cos y - \boldsymbol{i} \sin y = e^{-\boldsymbol{i}\,y}$$

et puisque $e^z = e^x e^{\boldsymbol{i}\,y}$, il vient $\overline{e^z} = \overline{e^x e^{\boldsymbol{i}\,y}} = e^x\, \overline{e^{\boldsymbol{i}\,y}} = e^x e^{-\boldsymbol{i}\,y} = e^{x-\boldsymbol{i}\,y} = e^{\overline{z}}$.
On a aussi $|e^{\boldsymbol{i}\,y}| = |\cos y + \boldsymbol{i} \sin y| = 1$, donc $|e^z| = \big|e^x e^{\boldsymbol{i}\,y}\big| = |e^x||e^{\boldsymbol{i}\,y}| = e^x$, car e^x est positif.
Par définition, $\operatorname{Arg}(e^z) = \operatorname{Arg}(e^{\boldsymbol{i}\,y})$ est égal à y modulo 2π.

3) Supposons que u est un nombre complexe tel que $e^u = 1$ et posons $u = a + b\,\boldsymbol{i}$, où a et b sont des nombres réels. On a $1 = |e^u| = e^a$, donc $a = 0$ car la fonction exponentielle réelle

est une bijection de $\mathbb{R}$ dans $]0,+\infty[$. On a aussi $\mathrm{Arg}(1)=0$ et $\mathrm{Arg}(e^u)=b$ modulo 2π, donc b est de la forme $2k\pi$, où $k\in\mathbb{Z}$. Il s'ensuit $u=b\boldsymbol{i}=2k\pi\boldsymbol{i}$. Réciproquement, pour tout entier k, on a $e^{2k\pi\boldsymbol{i}}=\cos(2k\pi)+\boldsymbol{i}\sin(2k\pi)=1$. Cela montre que les nombres complexes u tels que $e^u=1$ sont ceux de la forme $2k\pi\boldsymbol{i}$, où k est un entier relatif.
Soient maintenant des nombres complexes z et z'. D'après (i), on a l'équivalence $e^{z'}=e^z \iff e^{z'-z}=e^0=1$. D'après ce qui précède, cela équivaut à $z'-z=2k\pi\boldsymbol{i}$, où $k\in\mathbb{Z}$. ■

Proposition. *Si x et y sont des nombres réels, alors* $|e^{\boldsymbol{i}x}-e^{\boldsymbol{i}y}|=2|\sin\dfrac{x-y}{2}|$.

Démonstration. Nous avons $e^{\boldsymbol{i}x}-e^{\boldsymbol{i}y}=e^{\boldsymbol{i}y}(e^{\boldsymbol{i}(x-y)}-1)$ donc $|e^{\boldsymbol{i}x}-e^{\boldsymbol{i}y}|=|e^{\boldsymbol{i}y}||e^{\boldsymbol{i}(x-y)}-1|=|e^{\boldsymbol{i}(x-y)}-1|$ car $e^{\boldsymbol{i}y}$ est de module 1. Posons $\theta=x-y$. Par définition du module, il vient

$$\begin{aligned}|e^{\boldsymbol{i}\theta}-1|^2&=(e^{\boldsymbol{i}\theta}-1)(\overline{e^{\boldsymbol{i}\theta}-1})=(e^{\boldsymbol{i}\theta}-1)(e^{-\boldsymbol{i}\theta}-1)\\&=e^{\boldsymbol{i}\theta}e^{-\boldsymbol{i}\theta}-(e^{\boldsymbol{i}\theta}+e^{-\boldsymbol{i}\theta})+1\\&=1-2\,\mathcal{R}e(e^{\boldsymbol{i}\theta})+1=2-2\cos\theta\,.\end{aligned}$$

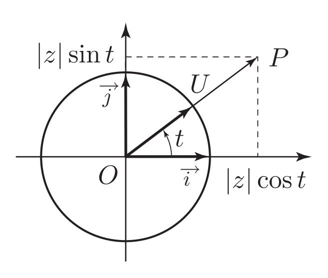
Puisque $1-\cos\theta=2\sin^2\dfrac{\theta}{2}$, on obtient $|e^{\boldsymbol{i}\theta}-1|^2=4\sin^2\dfrac{\theta}{2}$, d'où le résultat en prenant les racines carrées. ■

1.3 Utilisation géométrique des nombres complexes

Donnons-nous un repère orthonormé $(O;\vec{i},\vec{j})$ du plan euclidien.
Si z est un nombre complexe, le point de coordonnées (a,b) s'appelle le point d'*affixe* z (dans le repère). Le nombre z peut donc se représenter par un point du plan.
Si P' est le point de coordonnées (a',b'), l'affixe de P' est le nombre $z'=a'+b'\boldsymbol{i}$ et le vecteur $\overrightarrow{PP'}$ est représenté par le nombre complexe $z'-z=(a'-a)+(b'-b)\boldsymbol{i}$.

Distance. La distance PP' est égale à $\sqrt{(a'-a)^2+(b'-b)^2}=|z'-z|$. En particulier, la distance OP est égale au module de z.

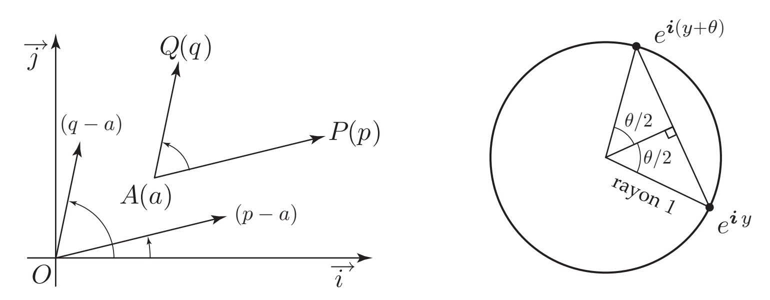
Angle polaire. Soit P un point du plan différent de l'origine. Si z est l'affixe de P, les coordonnées de P sont $(|z|\cos t,|z|\sin t)$, où $t=\mathrm{Arg}\,z$. Soit U le point de coordonnées $(\cos t,\sin t)$. L'angle de vecteurs $\widehat{\vec{i},\overrightarrow{OU}}$ a pour mesure t. On a $\overrightarrow{OP}=|z|\overrightarrow{OU}$ et $|z|>0$, donc $\widehat{\vec{i},\overrightarrow{OP}}=\widehat{\vec{i},\overrightarrow{OU}}$. L'angle $\widehat{\vec{i},\overrightarrow{OP}}$ s'appelle *l'angle polaire* de P. L'angle polaire d'un point d'affixe $z\neq 0$ a donc pour mesure $\mathrm{Arg}\,z$.



Coordonnées polaires. Puisqu'un nombre complexe non nul est déterminé par son module et son argument, tout point M du plan, différent de l'origine, est déterminé par la distance OM et l'angle polaire $\widehat{\vec{i},\overrightarrow{OM}}$. Soit $\theta\in[0,2\pi[$ la mesure de l'angle polaire et soit $r=OM$: les nombres r,θ s'appellent les *coordonnées polaires* du point M.

Angle de vecteurs. Soient A, P, Q des points du plan tels que $P \neq A$ et $Q \neq A$. Les vecteurs non nuls $\overrightarrow{AP}$ et $\overrightarrow{AQ}$ forment l'angle $\widehat{\overrightarrow{AP}, \overrightarrow{AQ}} = \left(\widehat{\vec{i}, \overrightarrow{AQ}}\right) - \left(\widehat{\vec{i}, \overrightarrow{AP}}\right)$.

Notons p l'affixe du point P et q l'affixe du point Q. Comme le vecteur $\overrightarrow{AP}$ est représenté par le nombre complexe $p - a$, l'angle $\widehat{\vec{i}, \overrightarrow{AP}}$ a pour mesure l'argument de $p - a$; de même, l'angle $\widehat{\vec{i}, \overrightarrow{AQ}}$ a pour mesure l'argument de $q - a$. Donc l'angle $\widehat{\overrightarrow{AP}, \overrightarrow{AQ}}$ a pour mesure $\mathrm{Arg}(q - a) - \mathrm{Arg}(p - a) = \mathrm{Arg}\left(\frac{q - a}{p - a}\right)$.



La figure de droite montre l'aspect géométrique de l'égalité $|e^{\boldsymbol{i}(y+\theta)} - e^{\boldsymbol{i}\,y}| = 2|\sin \frac{\theta}{2}|$, pour y et θ réels.

Produit scalaire. Si $\vec{U} = a\vec{i} + b\vec{j}$ et $\vec{U'} = a'\vec{i} + b'\vec{j}$ sont des vecteurs du plan, leur produit scalaire $\vec{U} \cdot \vec{U'}$ est le nombre $aa' + bb'$. On a $(\overline{a + b\,\boldsymbol{i}})(a' + b'\boldsymbol{i}) = (a - b\boldsymbol{i})(a' + b'\boldsymbol{i}) = aa' + bb' + (ab' - ba')\,\boldsymbol{i}$, donc

$$\vec{U} \cdot \vec{U'} = \mathscr{Re}(\bar{z}z'), \text{ où } z \text{ représente le vecteur } \vec{U} \text{ et } z' \text{ le vecteur } \vec{U'}.$$

Exemples de transformations du plan

Les glissements. Étudions la transformation $f : \mathbb{C} \to \mathbb{C}$ définie par $f(z) = \bar{z} + u$, où u est un nombre réel non nul.

Si M est un point d'affixe $z = x + y\boldsymbol{i}$, avec x et y des nombres réels, le point M' d'affixe $\bar{z} = x - y\boldsymbol{i}$ est symétrique de M par rapport à l'axe des abscisses. Soit M'' le point d'affixe $f(z) = \bar{z} + u$. Le vecteur $\overrightarrow{M'M''}$ a pour affixe $f(z) - \bar{z} = u$ et puisque u est réel, on a $\overrightarrow{M'M''} = u\vec{i}$. La transformation du plan définie par $M \mapsto M''$ est donc la composée $t \circ s$ de la translation t de vecteur $u\vec{i}$ et de la symétrie orthogonale s par rapport à l'axe Ox des abscisses : une telle transformation s'appelle un *glissement d'axe* Ox. On a

$$\begin{aligned} f \circ f(z) &= f(\bar{z} + u) = \overline{\bar{z} + u} + u \\ &= z + \bar{u} + u = z + 2u\,, \quad \text{car } u \text{ est réel.} \end{aligned}$$

Le point d'affixe $f \circ f(z)$ se déduit donc du point d'affixe z par la translation de vecteur $2u\vec{i}$.

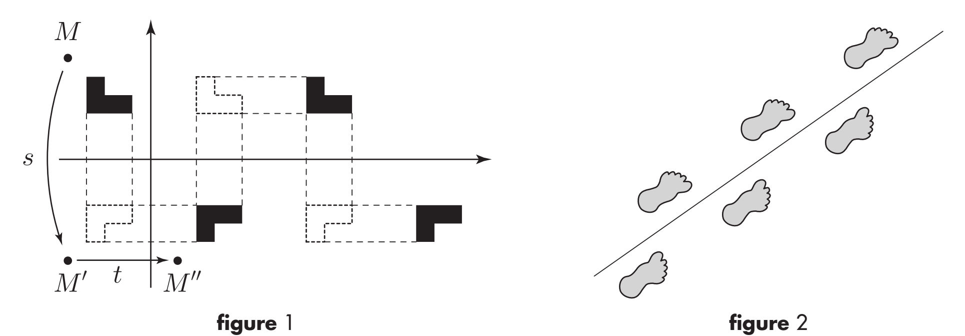


figure 1 **figure** 2

La figure 1 montre quelques itérés par le glissement f. Les glissements permettent de réaliser des frises, comme on le voit sur la figure 2 et dans l'exercice 13 page 33.

Les similitudes. Soient a un nombre complexe non nul, b un nombre complexe et $f : \mathbb{C} \to \mathbb{C}$ la fonction définie par $f(z) = az + b$.
Soit z un nombre complexe ; notons M le point d'affixe z et M' le point d'affixe $f(z)$.

- Si $a = 1$, on a $f(z) - z = b$, donc $\overrightarrow{MM'} = \overrightarrow{OB}$, où B est le point d'affixe b. Le vecteur $\overrightarrow{MM'}$ est fixe, donc la transformation du plan $M \mapsto M'$ est la translation de vecteur $\overrightarrow{OB}$. Si $b \neq 0$, il n'y a aucun point fixe ; si $b = 0$, alors $M' = M$ et la transformation est l'identité du plan.
- Supposons $a \neq 1$. Un nombre z est point fixe de f si et seulement si $az + b = z$, ce qui équivaut à $(1 - a)z = b$. L'unique point fixe de f est donc $\omega = b/(1 - a)$. En faisant avec des nombres complexes les mêmes calculs que dans l'exemple 2 page 16, on obtient l'égalité

  $(*) \qquad f(z) - \omega = a(z - \omega)$, pour tout $z \in \mathbb{C}$.

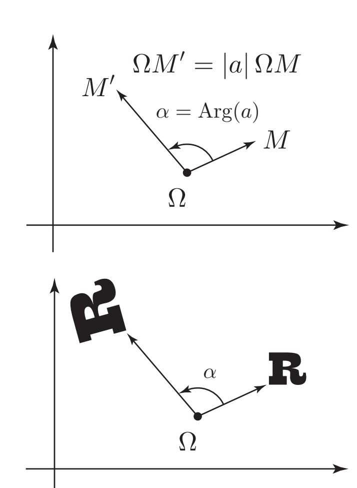
  Notons Ω le point d'affixe ω. La distance ΩM est égale à $|z - \omega|$, de même $\Omega M' = |f(z) - \omega|$. D'après $(*)$, on a $|f(z) - \omega| = |a(z - \omega)| = |a||z - \omega|$, donc

  $$\Omega M' = |a|\, \Omega M\,.$$

  L'angle de vecteurs $\widehat{\overrightarrow{\Omega M}, \overrightarrow{\Omega M'}}$ a pour mesure

  $$\operatorname{Arg}\left(\frac{f(z) - \omega}{z - \omega}\right) = \operatorname{Arg}(a)\,.$$

  Cela signifie que M' est l'image de M par la similitude de centre Ω, de rapport $|a|$ et d'angle $\operatorname{Arg}(a)$. Remarquons que si a est un nombre réel, f est une homothétie de rapport a. Si a n'est pas réel et si $|a| = 1$, alors f est une rotation de centre Ω.



Itération par une similitude. Exprimons les itérés $z_1 = az_0 + b$, $z_2 = az_1 + b$, … , $z_n = az_{n-1} + b$ d'un nombre z_0. D'après $(*)$, on a $z_1 - \omega = a(z_0 - \omega)$, $z_2 - \omega = a(z_1 - \omega)$ et en général $z_n - \omega = a(z_{n-1} - \omega)$. On en déduit

$$z_n - \omega = a^n(z_0 - \omega) \quad \text{pour tout entier } n \geqslant 1,$$

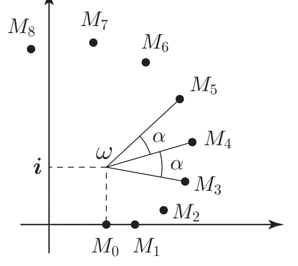
et comme $\omega = \dfrac{b}{1-a}$, il vient $z_n = a^n z_0 + \omega(1 - a^n) = a^n z_0 + b\dfrac{1-a^n}{1-a}$.

Exemple. La figure ci-contre montre les itérés du point M_0 d'affixe 1 par la similitude $z \mapsto \dfrac{2+\boldsymbol{i}}{2} z + \dfrac{1-\boldsymbol{i}}{2}$; le centre est le point d'affixe

$$\omega = \frac{1-\boldsymbol{i}}{2} \, \frac{1}{1-(1+\boldsymbol{i}/2)} = \frac{1-\boldsymbol{i}}{2} \, \frac{2}{-\boldsymbol{i}} = 1 + \boldsymbol{i},$$

le rapport est $\left|\dfrac{2+\boldsymbol{i}}{2}\right| = \dfrac{\sqrt{5}}{2}$ et l'angle est

$$\alpha = \operatorname{Arg}\left(\frac{2+\boldsymbol{i}}{2}\right) = \operatorname{Arg}\big(1 + (\boldsymbol{i}/2)\big).$$



Puisque la partie réelle et la partie imaginaire de $1 + (\boldsymbol{i}/2)$ sont positives, on a $0 < \alpha < \pi/2$; la tangente d'un argument est le rapport partie imaginaire sur partie réelle, donc $\operatorname{tg}(\alpha) = 1/2$, ce qui donne pour α environ 26,565 degrés.

Le groupe des similitudes

Si a est un nombre complexe non nul et si b est un nombre complexe quelconque, notons $f_{a,b} : \mathbb{C} \to \mathbb{C}$ l'application définie par $f_{a,b}(z) = az + b$. Pour tous nombres complexes z et z', on a l'équivalence

$$(*) \qquad z' = az + b \iff z = \frac{1}{a} z' - \frac{b}{a}$$

donc l'application $f_{a,b}$ est une transformation bijective de l'ensemble $\mathbb{C}$.

Soit T l'ensemble des applications $f_{a,b}$, où $a \neq 0$.

Nous allons montrer que l'ensemble T est un groupe de transformations de $\mathbb{C}$. Pour cela, vérifions les trois propriétés de la définition donnée page 28.

- On a $f_{1,0}(z) = z$ quel que soit z, donc $f_{1,0} = \mathrm{id}_{\mathbb{C}}$: la transformation $\mathrm{id}_{\mathbb{C}}$ appartient donc à T.
- Composons les transformations $f_{a,b}$ et $f_{c,d}$, où a et c sont non nuls : pour tout nombre complexe z, il vient

  $$\big(f_{a,b} \circ f_{c,d}\big)(z) = f_{a,b}(cz + d) = a(cz + d) + b = (ac)z + ad + b.$$

  Puisque ac n'est pas nul, la composée $f_{a,b} \circ f_{c,d} = f_{ac,ad+b}$ appartient à T.
- D'après l'équivalence $(*)$, la transformation réciproque de $f_{a,b}$ est $z' \mapsto \dfrac{1}{a} z' - \dfrac{b}{a} = f_{1/a,-b/a}(z')$, donc on a $f_{a,b}^{-1} = f_{1/a,-b/a}$. Cela montre que la transformation $f_{a,b}^{-1}$ appartient aussi à T.

L'ensemble T est donc un groupe de transformations de $\mathbb{C}$. Nous avons montré précédemment que les transformations du plan euclidien représentées par les applications $f_{a,b}$ sont les similitudes. L'ensemble des similitudes du plan euclidien est donc aussi un groupe de transformations.

2. Fonctions polynômes

2.1 Définitions et propriétés générales

Définitions

Un polynôme est une expression $P=p_0+p_1z+p_2z^2+\cdots+p_nz^n$, où $p_0,p_1,\ldots p_n$ sont des nombres réels ou complexes. Les nombres p_i sont les coefficients de P. Si le coefficient p_n n'est pas nul, l'entier n s'appelle le *degré* de P et se note $\deg P$. La fonction $P:\mathbb{C}\to\mathbb{C}$ définie par $P(z)=p_0+p_1z+p_2z^2+\cdots+p_nz^n$ est appelée *fonction polynôme*.

- La valeur $P(0)=p_0$ est le *terme constant* du polynôme.
- Une somme, un produit ou une composée de polynômes est un polynôme.
- Le produit d'un polynôme de degré p et d'un polynôme de degré q est un polynôme de degré $p+q$. Le polynôme nul n'a pas de degré.
- Il s'ensuit qu'un produit de polynômes n'est nul que si l'un des facteurs est nul.
- On peut aussi dériver un polynôme, en appliquant les règles valables pour une fonction d'une variable réelle. Ainsi, la dérivée du polynôme z^k est kz^{k-1}, où l'on convient (pour $k=1$) que l'on a $z^0=1$. Nous ferons toujours cette convention dans les calculs.

Définition

La *dérivée* du polynôme $P=p_0+p_1z+p_2z^2+\cdots+p_nz^n$ est le polynôme $P'=p_1+2p_2z+\cdots+np_nz^{n-1}$ si $n\geqslant 1$. Si $n=0$, le polynôme P est constant et l'on a $P'=0$.

On définit de proche en proche les dérivées successives de P : la dérivée seconde notée P'' et pour tout entier $k\geqslant 3$, la dérivée k-ième, notée $P^{(k)}$.

Exemples

- Si $P=z^n$, la dérivée k-ième de P est

$$P^{(k)}=\begin{cases} n(n-1)\cdots(n-k+1)z^{n-k}=\dfrac{n!}{k!}z^{n-k} & \text{si } 0\leqslant k\leqslant n \\ 0 & \text{si } k>n \end{cases}$$

- Supposons $P=p_0+p_1z+p_2z^2+p_3z^3$. On a $p_0=P(0)$ et $P'=p_1+2p_2z+3p_3z^2$, donc $p_1=P'(0)$. En dérivant à nouveau, il vient $P''=2p_2+3\times 2p_3z$, donc $2p_2=P''(0)$, puis $P^{(3)}=3\times 2p_3$, donc $6p_3=P^{(3)}(0)$.
- Plus généralement, pour un polynôme $P=p_0+p_1z+\cdots+p_nz^n$ de degré n, on a les relations $k!\,p_k=P^{(k)}(0)$ pour tout entier k tel que $0\leqslant k\leqslant n$. Les coefficients de

P s'expriment donc au moyen des dérivées en 0 par les formules : $p_k = \dfrac{P^{(k)}(0)}{k!}$.
On a ainsi l'égalité

$$P = P(0) + \frac{P'(0)}{1!} z + \frac{P''(0)}{2!} z^2 + \cdots + \frac{P^{(k)}(0)}{k!} z^k + \cdots + \frac{P^{(n)}(0)}{n!} z^n$$

Formule de Taylor pour les polynômes. *Si P est un polynôme de degré n, alors pour tout nombre complexe a, on a*

$$P = P(a) + \frac{P'(a)}{1!}(z-a) + \frac{P''(a)}{2!}(z-a)^2 + \cdots + \frac{P^{(k)}(a)}{k!}(z-a)^k + \cdots + \frac{P^{(n)}(a)}{n!}(z-a)^n.$$

Démonstration. Nous venons de voir que la formule est vraie si $a = 0$. Dans le cas général, définissons un polynôme Q en posant $Q(z) = P(z+a)$. D'après les règles de dérivation, on a $Q'(z) = P'(z+a)$, d'où $Q^{(k)}(z) = P^{(k)}(z+a)$ pour tout entier $k \geqslant 1$. On en déduit $Q(0) = P(a)$, $Q^{(k)}(0) = P^{(k)}(a)$ et il vient

$$\begin{aligned} Q(z) &= Q(0) + \frac{Q'(0)}{1!} z + \frac{Q''(0)}{2!} z^2 + \cdots + \frac{Q^{(k)}(0)}{k!} z^k + \cdots + \frac{Q^{(n)}(0)}{n!} z^n \\ &= P(a) + \frac{P'(a)}{1!} z + \frac{P''(a)}{2!} z^2 + \cdots + \frac{P^{(k)}(a)}{k!} z^k + \cdots + \frac{P^{(n)}(a)}{n!} z^n. \end{aligned}$$

Remplaçons z par $z - a$; puisque $Q(z-a) = P(z)$, on obtient la formule annoncée. ■

2.2 Racines d'un polynôme

Définition

Soit P un polynôme. Un nombre complexe a tel que $P(a) = 0$ s'appelle une *racine* de P. Si $P(a) = P'(a) = 0$, on dit que a est *racine multiple* de P.

Exemples

- Les nombres $\boldsymbol{i}$ et $-\boldsymbol{i}$ sont racines du polynôme $z^2 + 1$.
- Posons $j = \dfrac{-1}{2} + \dfrac{\sqrt{3}}{2}\boldsymbol{i}$. On a $j^2 = \dfrac{-1}{2} - \dfrac{\sqrt{3}}{2}\boldsymbol{i} = \bar{j}$, d'où $j + j^2 = j + \bar{j} = 2\mathcal{Re}(j) = -1$. Il vient donc la relation $1 + j + j^2 = 0$. Cela signifie que j est racine du polynôme $z^2 + z + 1$; l'autre racine est $\bar{j}$, car $1 + \bar{j} + \bar{j}^2 = 1 + \overline{j + j^2} = 1 + (-1) = 0$.

Proposition. *Soit P un polynôme à coefficients réels.*

- *Pour tout nombre complexe z, on a $\overline{P(z)} = P(\bar{z})$.*
- *Si a est racine de P, alors $\bar{a}$ est racine de P.*

Démonstration. Soit $P = p_0 + p_1 z + \cdots + p_n z^n$ un polynôme dont les coefficients $p_0, \ldots, p_n$ sont tous réels. On a

$$\begin{aligned} \overline{P(z)} &= \overline{p_0 + p_1 z + \cdots + p_n z^n} = \overline{p_0} + \overline{p_1 z} + \cdots + \overline{p_n z^n} \\ &= p_0 + p_1 \bar{z} + \cdots + p_n \overline{z^n} \quad \text{car } \overline{p_i z^i} = \overline{p_i}\,\overline{z^i} \text{ et } \overline{p_i} = p_i \\ &= p_0 + p_1 \bar{z} + \cdots + p_n \bar{z}^n \quad \text{car } \overline{z^i} = \bar{z}^i. \\ &= P(\bar{z}). \end{aligned}$$

Si a est racine de P, alors on a $P(a) = 0$, donc aussi $\overline{P(a)} = 0$ et par suite $P(\bar{a}) = 0$. ■

Proposition. *Soit P un polynôme.*

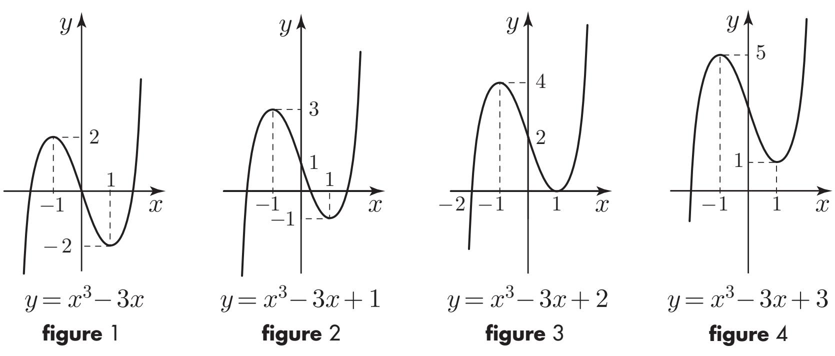
- *Un nombre a est racine de P si et seulement s'il existe un polynôme Q tel que $P=(z-a)Q$.*
- *Un nombre a est racine multiple de P si et seulement s'il existe un polynôme R tel que $P = (z-a)^2R$.*

Démonstration. Si Q existe, alors $P(a) = (a-a)Q(a) = 0$, donc a est racine de P. Réciproquement, supposons $P(a) = 0$. D'après la formule de Taylor, nous avons

$$\begin{aligned} P &= P(a) + P'(a)(z-a) + \frac{P''(a)}{2!}(z-a)^2 + \cdots + \frac{P^{(n)}(a)}{n!}(z-a)^n \\ &= (z-a)\left[P'(a) + \frac{P''(a)}{2!}(z-a) + \cdots + \frac{P^{(n)}(a)}{n!}(z-a)^{n-1}\right] \end{aligned}$$

expression qui est bien de la forme $(z-a)Q$, où Q est un polynôme.
Supposons que $P = (z-a)^2R$, où R est un polynôme. En dérivant, on obtient $P' = 2(z-a)R + (z-a)^2R'$, donc $P'(a) = 0$. Réciproquement, supposons $P(a) = P'(a) = 0$. Alors le premier terme de la formule de Taylor est $\dfrac{P''(a)}{2!}(z-a)^2$ et $(z-a)^2$ est en facteur dans l'expression de P. ■

Exemple. Posons $P=z^3-3z+a$, où a est un nombre réel. Les racines de $P'=3z^2-3$ sont ± 1 et l'on a $P(1) = -2+a$, $P(-1) = 2+a$: le polynôme P n'a donc une racine multiple que si $a = \pm 2$.
Pour étudier les racines de P, nous avons représenté ci-dessous les graphes des fonctions $x \mapsto x^3 - 3x + a$ lorsque a prend les valeurs 0, 1, 2 et 3.



$y = x^3 - 3x$ **figure** 1 — $y = x^3 - 3x + 1$ **figure** 2 — $y = x^3 - 3x + 2$ **figure** 3 — $y = x^3 - 3x + 3$ **figure** 4

- Si $a \in]-2,2[$, on voit que le graphe de la fonction $x \mapsto x^3 - 3x + a$ coupe l'axe des abscisses en trois points : le polynôme P a donc trois racines réelles (figures 1 et 2). En ces points, le graphe n'est pas tangent à l'axe des abscisses, ce qui traduit le fait que les racines sont simples, c'est-à-dire non multiples.
- Si $a = 2$, le graphe est tangent à l'axe au point d'abscisse 1 : cela correspond aux égalités $P(1) = P'(1) = 0$, donc 1 est racine double de P (figure 3). La factorisation

$P = z^3 - 3z + 2 = (z-1)^2(z+2)$ montre que l'autre racine est -2. De même, si $a = -2$, alors $P(-1) = P'(-1) = 0$ et -1 est racine double de P.

- Supposons $a > 2$. La fonction $x \mapsto P(x)$ ne coupe l'axe des abscisses qu'en un seul point $x_a < -2$: le polynôme P n'a qu'une racine réelle (figure 4). On a une factorisation $P = (z - x_a)(z^2 + pz + q)$, où p et q sont réels car P est à coefficients réels. Les autres racines de P sont donc celles du trinôme $z^2 + pz + q$: celles-ci sont nécessairement non réelles et conjuguées l'une de l'autre. Les trois racines de P sont simples.

Exemple. Soit n un entier au moins égal à 2 et soit $a \in \mathbb{C}$. Le polynôme $z^n - a^n$ a pour racine a, donc se factorise par $z - a$. Par exemple, on a les égalités

$$z^2 - a^2 = (z-a)(z+a) \qquad \text{et} \qquad z^3 - a^3 = (z-a)(z^2 + az + a^2)$$

Voici la formule générale pour factoriser $z^n - a^n$ par $z - a$:

$$(*) \qquad z^n - a^n = (z-a)(z^{n-1} + z^{n-2}a + \cdots + z^{n-k}a^{k-1} + \cdots + za^{n-2} + a^{n-1})$$

La formule est évidente si $a = 0$. Démontrons-la d'abord pour $a = 1$. Si z est un nombre complexe, on a

$$z(z^{n-1} + z^{n-2} + \cdots + z + 1) = z^n + z^{n-1} + \cdots + z$$
$$1(z^{n-1} + z^{n-2} + \cdots + z + 1) = \qquad z^{n-1} + \cdots + z + 1$$

donc en soustrayant $(z-1)(z^{n-1} + z^{n-2} + \cdots + z + 1) = z^n - 1$. Dans le cas d'un nombre complexe $a \neq 0$ quelconque, en remplaçant z par z/a dans l'égalité ci-dessus, on obtient $\left[\frac{z}{a} - 1\right]\left[\frac{z^{n-1}}{a^{n-1}} + \frac{z^{n-2}}{a^{n-2}} + \cdots + \frac{z}{a} + 1\right] = \frac{z^n}{a^n} - 1$, d'où l'égalité $(*)$ en multipliant chaque membre par a^n.

Remarque

L'égalité $(z-1)(1 + z + \cdots + z^{n-1}) = z^n - 1$ s'écrit encore

$$1 + z + \cdots + z^{n-1} = \frac{1 - z^n}{1 - z}\ , \quad \text{si } z \neq 1.$$

C'est la formule qui exprime la somme des n premiers termes de la progression géométrique de premier terme 1 et de raison $z \neq 1$.

Corollaire. *Un polynôme de degré n possède au plus n racines.*

Démonstration. Supposons que les nombres $a_1, \ldots, a_p$ sont des racines deux à deux différentes du polynôme P. D'après la proposition, il y a un polynôme Q_1 tel que $P = (z - a_1)Q_1$. On a $0 = P(a_2) = (a_2 - a_1)Q_1(a_2)$ et $a_2 - a_1 \neq 0$, donc $Q_1(a_2) = 0$. En appliquant la proposition à Q_1, il vient $Q_1 = (z - a_2)Q_2$, donc $P = (z - a_1)(z - a_2)Q_2$, où Q_2 est un polynôme. On a $Q_2(a_3) = 0$, car $(a_3 - a_1)(a_3 - a_2) \neq 0$. En continuant ainsi, on obtient une factorisation $P = (z - a_1)(z - a_2)\cdots(z - a_p)Q_p$. Puisque le degré d'un produit est la somme des degrés, on en déduit $\deg P = p + \deg Q_p \geqslant p$. ■

Un polynôme de degré n peut avoir moins de n racines : ainsi par exemple, le polynôme $(z-a)^4$ est de degré 4 et sa seule racine est a. Voici le résultat fondamental concernant les racines d'un polynôme.

Théorème de D'Alembert-Gauss. *Tout polynôme P de degré $n \geqslant 1$ s'écrit*

$$P = a(z-u_1)^{n_1}(z-u_2)^{n_2}\cdots(z-u_k)^{n_k}\ ,$$

avec $a \neq 0$, des nombres complexes $u_1,\ldots,u_k$ deux à deux différents et des entiers $n_1,\ldots,n_k$ au moins égaux à 1. On a $n_1+n_2+\cdots+n_k = n$ et les racines de P sont $u_1,\ldots,u_k$.

Ce théorème, que nous admettons, affirme notamment qu'un polynôme P non constant possède toujours des racines dans $\mathbb{C}$ et que P est un produit de polynômes $z-u_i$ de degré 1, chacun des facteurs pouvant être répété plusieurs fois. L'entier n_i s'appelle la *multiplicité* de la racine u_i. Si $n_i=1$, on dit que u_i est *racine simple* de P.

Racines n-ièmes de l'unité. Soit n un entier au moins égal à 2.
Pour tout entier k tel que $0 \leqslant k \leqslant n-1$, posons $u_k = e^{(2k\pi/n)\boldsymbol{i}}$. D'après la formule de Moivre, on a

$$(u_k)^n = \big[\cos(2k\pi/n) + \boldsymbol{i}\sin(2k\pi/n)\big]^n = \cos(2k\pi) + \boldsymbol{i}\sin(2k\pi) = 1\,,$$

donc les n nombres complexes $u_0,\ldots,u_{n-1}$ sont racines du polynôme z^n-1. Le polynôme z^n-1 étant de degré n, il n'a pas d'autre racine, d'où la factorisation

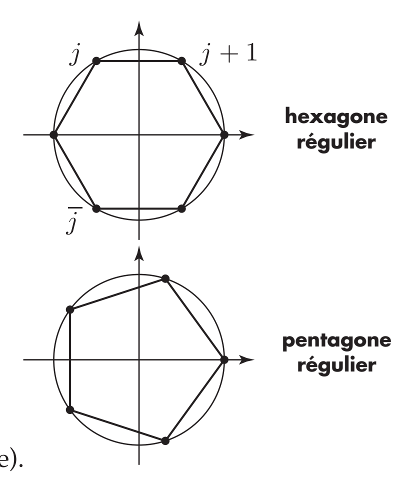
$$z^n - 1 = (z-1)\left(z - e^{\frac{2\pi}{n}\boldsymbol{i}}\right)\left(z - e^{2\frac{2\pi}{n}\boldsymbol{i}}\right)\cdots\left(z - e^{(n-1)\frac{2\pi}{n}\boldsymbol{i}}\right)\,.$$

Les nombres $u_k = e^{(2k\pi/n)\boldsymbol{i}}$ s'appellent les *racines n-ièmes de l'unité*. Remarquons que l'on a l'égalité $u_k = (u_1)^k$, avec la convention habituelle $(u_1)^0 = 1$.

Exemples

- Les racines carrées de l'unité sont 1 et -1. Les racines quatrièmes de l'unité sont les solutions de l'équation $z^4-1=(z^2-1)(z^2+1)=0$: ce sont les nombres 1, -1, $\boldsymbol{i}$, $-\boldsymbol{i}$.
- Les racines cubiques de l'unité sont les racines du polynôme $z^3-1=(z-1)(z^2+z+1)$, c'est-à-dire nombres 1, $j = e^{2\boldsymbol{i}\pi/3} = \dfrac{-1}{2} + \boldsymbol{i}\,\dfrac{\sqrt{3}}{2}$ et $j^2 = e^{4\boldsymbol{i}\pi/3} = \bar{j} = \dfrac{-1}{2} - \boldsymbol{i}\,\dfrac{\sqrt{3}}{2}$.

Puisqu'on a $|u_k|=1$, les points $P_0,\ldots,P_{n-1}$ d'affixes $u_0,\ldots,u_{n-1}$ sont sur le cercle de centre O et de rayon 1. De plus, pour $k=1,2,\ldots,n-1$, on a $u_k=(u_1)^k=u_1u_{k-1}$, donc l'argument de u_k/u_{k-1} vaut toujours $\mathrm{Arg}(u_1)=2\pi/n$. Cela veut dire que les angles $\widehat{\overrightarrow{OP_0},\overrightarrow{OP_1}}$, $\widehat{\overrightarrow{OP_1},\overrightarrow{OP_2}}$, etc, ont tous la même mesure $2\pi/n$, autrement dit les points $P_0,\ldots,P_{n-1}$ partagent le cercle en n arcs égaux. Les points $P_0,\ldots,P_{n-1}$ sont donc les sommets d'un polygone régulier à n cotés centré à l'origine. Les figures montrent le polygone obtenu avec les racines 6-ièmes de l'unité (hexagone) et celui correspondant aux racines 5-ièmes (pentagone).



Racines n-ièmes d'un nombre complexe. Soit a un nombre complexe non nul. Le module $|a|$ est réel et strictement positif. Puisque la fonction $x \mapsto x^n$ est une bijection de $[0,+\infty[$ dans $[0,+\infty[$, le nombre réel $\sqrt[n]{|a|}$ existe. Posons $\alpha = \operatorname{Arg} a$ et $z_k = \sqrt[n]{|a|}\, e^{\frac{\alpha}{n}\boldsymbol{i}}\, e^{\frac{2k\pi}{n}\boldsymbol{i}}$ pour k entier tel que $0 \leqslant k \leqslant n-1$. Les nombres $e^{\frac{2k\pi}{n}\boldsymbol{i}}$ étant racines n-ièmes de l'unité, on a

$$(z_k)^n = \left(\sqrt[n]{|a|}\right)^n \left(e^{\frac{\alpha}{n}\boldsymbol{i}}\right)^n = |a|e^{\alpha \boldsymbol{i}} = a$$

donc $z_0, \ldots, z_{n-1}$ sont les n racines du polynôme $z^n - a$. On en déduit la factorisation

$$z^n - a = (z - z_0) \cdots (z - z_{n-1})\,.$$

Proposition. *Soit a un nombre complexe non nul. Les racines n-ièmes de a sont les nombres complexes de module $\sqrt[n]{|a|}$ et d'argument $\dfrac{\operatorname{Arg} a}{n} + \dfrac{2k\pi}{n}$, où $k = 0, 1, \ldots, n-1$.*

Un nombre non nul, réel ou complexe, possède donc n racines n-ièmes. Remarquons que l'on a $z_0 = \sqrt[n]{|a|}\, e^{\frac{\alpha}{n}\boldsymbol{i}}$ et donc $z_k = z_0\, e^{\frac{2k\pi}{n}\boldsymbol{i}}$: les racines n-ièmes de a s'obtiennent en multipliant z_0 par les racines n-ièmes de l'unité.
Les racines carrées d'un nombre complexe non nul sont deux nombres opposés.

2.3 Calcul des valeurs d'une fonction polynôme

Soit $P = p_n z^n + p_{n-1} z^{n-1} + \cdots + p_1 z + p_0$ un polynôme non nul et soit u un nombre complexe. Pour obtenir la valeur $P(u)$, on peut calculer de proche en proche les puissances de u, retenir ces valeurs et en faire la combinaison selon les coefficients du polynôme. Voici un algorithme plus rapide : pour un polynôme $p_3 z^3 + p_2 z^2 + p_1 z + p_0$ par exemple, on calcule successivement les nombres

$$a_2 = p_3 u + p_2 \quad , \quad a_1 = a_2 u + p_1 \quad \text{et} \quad a_0 = a_1 u + p_0$$

de sorte qu'on a

$$\begin{aligned} a_0 &= a_1 u + p_0 = (a_2 u + p_1)u + p_0 \\ &= a_2 u^2 + p_1 u + p_0 \\ &= (p_3 u + p_2)u^2 + p_1 u + p_0 \\ &= p_3 u^3 + p_2 u^2 + p_1 u + p_0 = P(u)\,. \end{aligned}$$

Voici l'algorithme général, appelé *méthode de Hörner* :

initialisation : ($n \leftarrow \deg P$) (pour i de 0 à $\deg P$, $p[i] \leftarrow$ coefficient de z^i) ($a \leftarrow p[n]$)
  ($u \leftarrow$ un nombre complexe)

boucle : tant que $n > 0$, faire
  $a \leftarrow au + p[n-1]$
  $n \leftarrow n - 1$

fin : la valeur de a est $P(u)$.

2.4 Division euclidienne des polynômes

Nous avons montré qu'un nombre complexe a est racine d'un polynôme A si et seulement si A est multiple du polynôme $z - a$. Plus généralement, si A et B sont des polynômes, on peut se demander si A est multiple de B, c'est-à-dire s'il existe un polynôme Q tel que $A = BQ$ et si oui, comment calculer le quotient Q.
Étant donnés des polynômes A et B, le polynôme A n'est en général pas multiple de B, mais on peut trouver, parmi les multiples de B, celui qui diffère de A par un polynôme de plus petit degré possible. En appelant BQ ce multiple, on aura précisément

$$\deg(BQ) = \deg A \quad \text{et} \quad \deg(A - BQ) < \deg B \text{ si } A \neq BQ.$$

Exactement comme pour les entiers (voir page 4), on dispose sur les polynômes d'une division avec reste : le reste de la division du polynôme A par le polynôme B doit être, pour ce qui est du degré, « plus petit » que B.

Théorème. *Soit A un polynôme et soit B un polynôme non nul. Il existe des polynômes Q et R uniques tels que $A = BQ + R$ et $(R = 0$ ou $\deg R < \deg B)$. Le polynôme R s'appelle le reste de la division de A par B et Q est le quotient. Le polynôme A est multiple de B si et seulement si le reste de la division de A par B est nul.*

Calculer le quotient Q et le reste R, c'est effectuer *la division euclidienne* de A par B. On a $A = BQ$ si et seulement si $R = 0$. Cela veut dire que le polynôme A est multiple du polynôme B si et seulement si le reste de la division de A par B est nul. La division est donc un outil pour savoir si un polynôme se factorise par un autre.
Montrons sur des exemples comment se pratique la division.

Exemple 1. Posons $A = 2z^4 + z^3 - 3z + 5$ et $B = z^2 + 1$. En ajoutant (ou en retranchant) à A un multiple de B, on cherche à transformer A en un polynôme A_1 de degré plus petit que 4. Visiblement, si l'on retranche à A le polynôme $2z^2B = 2z^4 + 2z^2$, le terme $2z^4$ disparaît de A et l'on obtient le polynôme

$$A_1 = A - 2z^2B = z^3 - 2z^2 - 3z + 5$$

qui est de degré $3 < \deg A$. Continuons ainsi : en retranchant à A_1 le polynôme $zB = z^3 + z$, on obtient

$$A_2 = A_1 - zB = -2z^2 - 4z + 5.$$

On forme enfin le polynôme

$$R = A_2 + 2B = -2z^2 - 4z + 5 + (2z^2 + 2) = -4z + 7$$

qui est de degré $1 < \deg B$. Par conséquent, le reste de la division de A par B est $R = -4z + 7$. Pour calculer le quotient, exprimons A au moyen de A_1 , puis de A_2 et R : il vient

$$A = 2z^2B + A_1 = 2z^2B + zB + A_2 = 2z^2B + zB - 2B + R = (2z^2 + z - 2)B + R.$$

Le quotient est $Q = 2z^2 + z - 2$ et l'on a bien l'égalité de division $A = BQ + R$.

Dans la pratique, on conduit les calculs comme pour une division ordinaire entre nombres, en écrivant au fur et à mesure les quotients partiels sous le diviseur.

Exemple 2. Calculons le quotient et le reste de la division de $z^5+z^4+z^3+8z^2+9z+8$ par z^3+8.
On écrit

$$\begin{array}{rrrrrrrrrrr|l}
z^5 & + & z^4 & + & z^3 & + & 8z^2 & + & 9z & + & 8 & z^3+8 \\ \cline{12-12}
-z^5 & & & & & - & 8z^2 & & & & & z^2+z+1 \\
 & & z^4 & + & z^3 & & & + & 9z & + & 8 & \\
 & & -z^4 & & & & & - & 8z & & & \\
 & & & & z^3 & & & + & z & + & 8 & \\
 & & & & -z^3 & & & & & - & 8 & \\
 & & & & & & & & z & & &
\end{array}$$

Le quotient z^2+z+1 apparaît sous le diviseur, le reste est z et l'on a l'égalité de division

$$z^5+z^4+z^3+8z^2+9z+8=(z^3+8)(z^2+z+1)+z\,.$$

Exemple 3. Factorisons le polynôme $P=-z^4+2z^3-5z^2+8z-4$ en produit de facteurs de degré 1.
On a $P(1)=0$, donc 1 est racine de P. Par suite, P est multiple de $z-1$. En effectuant la division de P par $z-1$, on obtient

$$\begin{aligned} P &= (z-1)(-z^3+z^2-4z+4) = -(z-1)(z^3-z^2+4z-4) \\ &= -(z-1)\big[z^2(z-1)+4(z-1)\big] = -(z-1)^2(z^2+4)\,. \end{aligned}$$

Les racines de z^2+4 sont $2\boldsymbol{i}$ et $-2\boldsymbol{i}$, donc $P=-(z-1)^2(z-2\boldsymbol{i})(z+2\boldsymbol{i})$.

Division par $z-a$. *Si P est un polynôme et a un nombre, le reste de la division de P par $z-a$ est $P(a)$.*

En effet, d'après la formule de Taylor pour les polynômes, il existe un polynôme Q tel que $P=P(a)+(z-a)Q$, ou encore $P=(z-a)Q+P(a)$.

Polynômes étrangers

Définition
Si des polynômes non nuls A et B n'ont pas de racine commune (réelle ou complexe), ont dit qu'ils sont *étrangers*.

Exemple. Si a et b sont des nombres complexes différents, les polynômes $(z-a)^n$ et $(z-b)^p$ sont étrangers, quels que soient les entiers positifs n et p.

Proposition. *Des polynômes A et B sont étrangers si et seulement s'il existe des polynômes U et V tels que $AU+BV=1$.*

Démonstration. Une racine commune à A et B est aussi racine de $AU+BV$: si $AU+BV=1$, les polynômes A et B sont donc étrangers. Réciproquement, supposons A et B étrangers, $\deg B \leqslant \deg A$ et raisonnons par récurrence sur le degré de B. Si $\deg B = 0$, alors B est une constante b non nulle et l'on a $0A + (1/b)B = 1$. Supposons $\deg B \geqslant 1$ et effectuons la division euclidienne de A par B : $A = BQ + R$, où le reste est R. D'après cette relation, toute racine commune à B et R est racine commune à A et B, donc B et R n'ont pas de racine commune. De plus, R n'est pas nul, sinon A serait multiple de B et les racines de B seraient racines de A. Ainsi B et R sont étrangers. Puisqu'on a $\deg R < \deg B$, l'hypothèse de récurrence assure qu'il existe des polynômes U et V tels que $BU + RV = 1$. Il vient alors $AV = BQV + RV = BQV + 1 - BU = B(QV - U) + 1$, d'où $AV + B(U - QV) = 1$. ■

2.5 Rappels sur l'équation du second degré

Une équation du second degré est une équation de la forme

$$z^2 + pz + q = 0, \quad \text{où } p \text{ et } q \text{ sont des nombres réels ou complexes.}$$

Pour résoudre l'équation, on remarque que $z^2 + pz$ est le début du développement de $\left[z + (p/2)\right]^2$. Précisément, on a $z^2 + pz = \left[z + (p/2)\right]^2 - p^2/4$, d'où

$$z^2 + pz + q = \left[z + (p/2)\right]^2 - p^2/4 + q = \left[z + (p/2)\right]^2 - \frac{p^2 - 4q}{4}\,.$$

Posons $\Delta = p^2 - 4q$: c'est le *discriminant* du polynôme $z^2 + pz + q$.

Premier cas : $\Delta \neq 0$. Il existe alors deux nombres complexes δ et $-\delta$ tels que $\delta^2 = (-\delta)^2 = \Delta$. On a

$$z^2 + pz + q = \left(z + \frac{p}{2}\right)^2 - \left(\frac{\delta}{2}\right)^2 = \left(z + \frac{p}{2} - \frac{\delta}{2}\right)\left(z + \frac{p}{2} + \frac{\delta}{2}\right)$$

Les racines de z^2+pz+q sont donc les deux nombres $z_1 = -\frac{p}{2} + \frac{\delta}{2}$ et $z_2 = -\frac{p}{2} - \frac{\delta}{2}$.

Second cas : $\Delta = 0$. L'équation s'écrit $\left[z + (p/2)\right]^2 = 0$. Il n'y a qu'une solution : $z_1 = -p/2 = z_2$.

Dans tous les cas, on a la factorisation $z^2 + pz + q = (z - z_1)(z - z_2)$. Puisque $(z - z_1)(z - z_2) = z^2 - (z_1 + z_2)z + z_1 z_2$, on en déduit les égalités

$$p = -(z_1 + z_2) \quad \text{et} \quad q = z_1 z_2$$

qui expriment la somme et le produit des racines au moyen des coefficients de l'équation.

Cas où les coefficients sont réels. Supposons que les coefficients p et q sont des nombres réels. Le discriminant $\Delta = p^2 - 4q$ de l'équation est alors un nombre réel. L'égalité

$$x^2 + px + q = \left[x + (p/2)\right]^2 - \Delta/4$$

montre que lorsque x décrit l'ensemble des nombres réels, le minimum de x^2+px+q est obtenu pour la valeur $-p/2$ de la variable et que ce minimum vaut $-\Delta/4$.

Il est facile de déterminer le signe de $x^2 + px + q$:

- Si $\Delta < 0$, on a $x^2 + px + q = \left[x + (p/2)\right]^2 - \Delta/4 \geqslant -\Delta/4 > 0$ pour tout x réel.
- Si $\Delta=0$, on a $x^2+px+q=\left[x+(p/2)\right]^2 \geqslant 0$, la valeur 0 étant obtenue lorsque $x=-p/2$.
- Supposons $\Delta > 0$. L'équation a deux racines réelles x_1, x_2 et l'on a la factorisation $x^2 + px + q = (x - x_1)(x - x_2)$. Si x est strictement compris entre x_1 et x_2, alors $x^2+px+q<0$. Si x est en dehors de l'intervalle d'extrémités x_1,x_2, alors $x^2+px+q>0$. Le produit des racines est q : si $q < 0$, les racines sont de signes contraires ; si $q > 0$, les racines sont de même signe, celui de leur somme $-p$.

Exemple. On dispose d'une peinture dont deux composants a et b s'oxydent lentement à l'air. Après une série de mesures, on constate que si a et b sont présents en quantité x et y, alors au bout d'un an, les quantités sont

$$x' = y - kx \quad \text{et} \quad y' = ky - x,$$

où k est un paramètre positif qu'on peut faire varier en incorporant des produits additifs. Afin de stabiliser la peinture, on souhaite que la proportion entre les deux substances oxydables a et b ne varie pas. Cela est-il possible ?
La proportion initiale entre les quantités de a et de b est le rapport x/y. On doit donc chercher s'il existe des valeurs x et y telles que $x/y = x'/y'$.
Posons $t=x/y$ et $t'=x'/y'$. On a $x'=y-kx=y(1-kt)$ et $y'=ky-x=y(k-t)$, donc $t' = \dfrac{1-kt}{k-t}$. L'égalité $t = t'$ s'écrit $t = \dfrac{1-kt}{k-t}$, c'est-à-dire $t(k-t) = 1 - kt$ ou encore

$$(*) \qquad t^2 - 2kt + 1 = 0\,.$$

Cette équation du second degré a pour discriminant $\Delta = 4(k^2 - 1)$.

- Si $0 \leqslant k < 1$, l'équation $(*)$ n'a pas de solution réelle : pour ces valeurs du paramètre k, la proportion entre les produits a et b ne reste pas constante.
- Supposons $k \geqslant 1$. L'équation a pour solutions $t_1=k-\sqrt{k^2-1}$ et $t_2=k+\sqrt{k^2-1}$. Les nombres t_1 et t_2 sont positifs, car leur produit est 1 et leur somme $2k>0$. Puisque $y'=y(k-t)$, le nombre t doit vérifier l'inégalité $k-t>0$. On a $k-t_1=\sqrt{k^2-1}\geqslant 0$ et $k - t_2 = -\sqrt{k^2-1} \leqslant 0$, donc la seule possibilité est $k > 1$ et $t = t_1$.
- Supposons $k>1$ et posons $t=k-\sqrt{k^2-1}$. En partant de quantités positives x et y vérifiant $x=ty$, on obtient $y'=y(k-t)>0$ et aussi $x'=ty'>0$: au cours de la première année, la proportion entre a et b est restée fixe : nécessairement, elle le restera ensuite.

Comment les quantités x et y ont-elles varié en un an ? On a

$$\begin{aligned} y' &= y(k-t) = y\sqrt{k^2-1} \\ x' &= ty' = ty\sqrt{k^2-1} = x\sqrt{k^2-1} \end{aligned}$$

donc les quantités x et y ont été multipliées par le même facteur $\sqrt{k^2-1}$.
Puisque les valeurs x et y n'ont pu que diminuer au cours du temps, c'est que l'on a $\sqrt{k^2-1} < 1$, donc $1 < k < \sqrt{2}$.

Exercices

@ **1. Les transformations homographiques.** On ajoute à l'ensemble $\mathbb{C}$ des nombres complexes un élément ∞ qui n'est pas un nombre et l'on note $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ l'ensemble obtenu. Soient u, v, c, d des nombres complexes tels que $c \neq 0$ et $v \neq 0$ et soit $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ l'application définie en posant

$$f(z) = u + \frac{v}{cz+d} \text{ si } z \in \mathbb{C} \setminus \{-d/c\}\ ,\ \ f(-d/c) = \infty \text{ et } f(\infty) = u.$$

a) Montrer que f est une transformation bijective de $\overline{\mathbb{C}}$.

b) Montrer que l'application f possède deux points fixes si et seulement si $(d+uc)^2 + 4cv \neq 0$.

On suppose désormais que f possède deux points fixes p et q et l'on définit l'application $\varphi : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ en posant

$$\varphi(z) = \frac{z-p}{z-q} \text{ si } z \in \mathbb{C} \setminus \{q\}\ ,\ \ \varphi(q) = \infty \text{ et } \varphi(\infty) = 1.$$

c) Montrer que φ est une bijection.

d) Transportons f par le changement de référentiel φ : on obtient une transformation g de $\overline{\mathbb{C}}$. Montrer qu'en posant $K = \dfrac{cq+d}{cp+d}$, on a

$$g(\infty) = \infty \text{ et } g(z) = Kz \text{ pour tout } z \in \mathbb{C}.$$

@ **2. Une itération homographique.** Soit $I =]1/2, +\infty[$. Pour tout nombre $x \in I$, on pose $f(x) = \dfrac{4x-1}{2x+1}$.

a) Déterminer des nombres u et v tels que $f(x) = u + \dfrac{v}{2x+1}$ pour tout $x \in I$.

b) Montrer que f définit une transformation de l'intervalle I et déterminer le point fixe de cette transformation.

c) Montrer que pour tout $x \in I$, on a $\dfrac{f(x)-1}{f(x)-1/2} = \dfrac{2}{3}\,\dfrac{x-1}{x-1/2}$.

d) Soit $x_0 \in I$ et soient $x_1, \ldots, x_n, \ldots$ les itérés de x_0 par l'application f. Montrer que pour tout entier $n \geqslant 1$, on a $\dfrac{x_n-1}{x_n-1/2} = \left(\dfrac{2}{3}\right)^n \dfrac{x_0-1}{x_0-1/2}$.

e) Quelle est la limite de x_n quand n tend vers l'infini ?

Dans les deux exercices suivants, on pourra utiliser la formule du binôme de Newton qui figure page 61.

3. Utilisation de la formule de Moivre

a) Exprimer $\cos 2a$ au moyen de $\cos a$ et de $\sin a$. Faire de même pour $\sin 2a$.

b) Montrer que $\tan 2a = \dfrac{2\tan a}{1-\tan^2 a}$ et $\tan 3a = \dfrac{3\tan a - \tan^3 a}{1 - 3\tan^2 a}$.

c) Démontrer les égalités $\cos 3a = 4\cos^3 a - 3\cos a$ et $\sin 3a = 3\sin a - 4\sin^3 a$.

@ **4. Mesures du pentagone régulier**

a) Démontrer l'identité $\cos 5a = \cos^5 a - 10\cos^3 a \sin^2 a + 5\cos a \sin^4 a$ en utilisant la formule de Moivre. En déduire que les nombres $\cos(2\pi/5)$ et $\cos(4\pi/5)$ sont racines du polynôme $P = 16z^5 - 20z^3 + 5z - 1$.

b) Montrer que $P = (z-1)(4z^2+2z-1)^2$. En déduire $\cos(2\pi/5) = \frac{\sqrt{5}-1}{4}$.

c) Montrer que le coté du pentagone régulier inscrit dans un cercle de rayon R a pour longueur $R\sqrt{\dfrac{5-\sqrt{5}}{2}}$. Pour une construction géométrique du pentagone, voir à la fin des exercices.

d) Montrer que $\cos(\pi/5) = \frac{\sqrt{5}+1}{4}$ (utiliser la valeur connue de $\cos(2\pi/5)$).

@ **5. a)** Soient a, b, c des nombres complexes de module 1 tels que $a+b+c=0$.

(i) Montrer que l'on a $1 + 2\,\mathscr{Re}(a\bar{b}) = 0$.

(ii) En déduire que l'argument de b/a est égal à $2\pi/3$ ou à $4\pi/3$.

b) Soient $\vec{u}$, $\vec{v}$ et $\vec{w}$ des vecteurs du plan euclidien. On suppose que ces vecteurs sont de même module non nul et que leur somme est nulle. Montrer qu'entre deux de ces vecteurs, l'angle mesure 120 degrés.

6. Cet exercice est une application du précédent. On considère le dispositif mécanique ci-contre formé de trois poulies et de fils où sont suspendues des masses égales.

a) Montrer qu'au point d'équilibre E, les tensions sur tous les fils ont le même module.

b) En déduire l'angle au point E entre les directions des fils.

c) On rappelle que si P, Q, R sont trois points d'un cercle de centre O, « l'angle inscrit » $\widehat{PQ,PR}$ vaut, à 180 degrés près, la moitié de « l'angle au centre » $\widehat{OQ,OR}$. En déduire une construction géométrique du point d'équilibre E.

7. Pavage hexagonal. On se place dans le plan euclidien muni d'un repère orthonormé. Soit P l'hexagone régulier inscrit dans le cercle de rayon 1 centré à l'origine et ayant deux cotés parallèles à l'axe Ox.

a) Quelles sont les affixes des sommets de P ? Quelle est la longueur du coté de P ?

b) Soit P' l'image de P par la translation de vecteur d'affixe 3 et soit P'' l'image de P par la translation de vecteur d'affixe $u = 1 + e^{i\pi/3}$. Montrer que P'' a un coté en commun avec chacun des hexagones P et P'.

c) Pour tous entiers relatifs k et n, notons $P_{k,n}$ l'image de P par la translation du vecteur d'affixe $k\boldsymbol{i}+nu$.

(i) Quels sont les couples (k,n) correspondant à des hexagones $P_{k,n}$ ayant avec P un coté en commun ?

(ii) Montrer que les intérieurs des hexagones $P_{k,n}$ forment un pavage du plan. Quelle est la distance entre les centres de deux hexagones ayant un coté en commun ?

@ **8. Expression d'une symétrie à l'aide des nombres complexes.** On se place dans le plan euclidien muni d'un repère orthonormé. Soit D une droite passant par l'origine O et soit a l'affixe d'un vecteur directeur de D. Pour tout $z \in \mathbb{C}$, on pose $s(z) = a\bar{z}/\bar{a}$ et si M est un point d'affixe z, on note M' le point d'affixe $s(z)$.

a) Montrer que si $M \in D$, alors $M' = M$.

b) Montrer que le vecteur d'affixe $a\,\boldsymbol{i}$ est orthogonal à D.

c) Montrer que si le vecteur $\overrightarrow{OM}$ est orthogonal à D, alors $\overrightarrow{OM'}$ est orthogonal à D.

d) En déduire que M' est le symétrique de M par rapport à D.

@ **9. Polynôme de degré 3 ayant une racine multiple.** Soit le polynôme $P = z^3 + pz + q$, où p et q sont des nombres réels ou complexes.

a) On suppose que P a une racine multiple a. Démontrer les égalités $a^2 = -p/3$ et $a(-p/3) + ap + q = 0$. En déduire que l'on a $4p^3 + 27q^2 = 0$.

b) Supposons $4p^3 + 27q^2 = 0$.

(i) Soit a une racine carrée de $-p/3$. Montrer que $q/2 = \pm a^3$ et que $P(a) = q - 2a^3$.

(ii) En déduire que a ou $-a$ est racine multiple de P.

10. Factorisations de polynômes

a) Soit le polynôme $P = 4z^4 - 12z^3 + 17z^2 - 24z + 18$. Effectuer la division de P par le polynôme $z^2 + 2$. En déduire les racines de P.

b) Montrer que 1 et -1 sont racines multiples du polynôme $Q = 3z^5 + z^4 - 6z^3 - 2z^2 + 3z + 1$. Trouver toutes les racines de Q.

c) Trouver un polynôme R de degré inférieur ou égal à 2 tel que $z^4 - 5z^3 + 5z^2 + 3z + 7 - R$ soit multiple du polynôme $(z-3)^2$.

11. Soient p et q des nombres réels. Notons z_1 et z_2 les solutions de l'équation $z^2 + pz + q = 0$. Montrer que z_1 et z_2 sont de parties réelles strictement négatives si et seulement si p et q sont strictement positifs (on distinguera le cas où les racines sont réelles du cas où elles ne le sont pas).

@ **12. Fonctions du second degré à deux variables.** Soient p et q des nombres réels et soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie par $f(x,y) = x^2 + 2pxy + qy^2$.

a) On suppose $p^2 < q$. Montrer que $f(x,y)$ est strictement positif si x ou si y n'est pas nul. Quelle est la valeur minimum de la fonction f ?

b) On suppose $p^2 > q$. Montrer qu'il y a deux nombres réels u et v différents tels que $f(x,y) = (x+uy)(x+vy)$ pour tout (x,y). En déduire que $f(x,y)$ ne garde pas un signe fixe.

c) On suppose $f(x,y) = x^2 - xy - 2y^2$.

(i) Décomposer $f(x,y)$ en produit de facteurs comme en **(b)**, en calculant les nombres u et v.

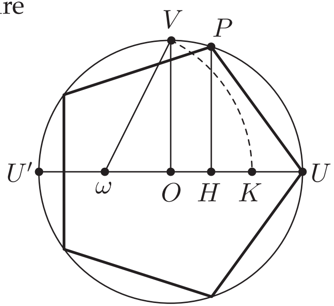
(ii) Définir un repère $(O;\vec{i'},\vec{j'})$ du plan euclidien $\mathbb{R}^2$ dont les axes ont pour équation $x+uy=0$ et $x+vy=0$. Quelle est l'équation d'une ligne de niveau de f dans ce nouveau repère ?

(iii) Dessiner sur un même dessin les lignes de niveau -1, 0, 1 et 2 de la fonction f.

@13. Effet multiplicateur en Économie. Quand des acteurs économiques (particuliers, collectivités ou entreprises) disposent d'une somme d'argent, ils en dépensent une partie auprès des autres et épargnent (ou capitalisent) le reste. Supposons qu'une politique de baisse d'impôts ou le financement de grands travaux, par exemple, mette à disposition des acteurs économiques une somme d'argent S. Supposons aussi que lorsqu'ils disposent d'une somme d'argent, les acteurs en redépensent tous la même fraction c, où $0 < c < 1$; le nombre c s'appelle *la proportion marginale à consommer*. Notons S_n le cumul des sommes utilisées en consommation après n transactions. Calculer S_n et montrer que $\lim\limits_{n\to+\infty} S_n = \frac{S}{1-c}$. Le coefficient $k = \frac{1}{1-c}$ s'appelle *le coefficient multiplicateur*. Combien vaut k lorsque les acteurs redépensent 75% de la somme à disposition ? Quel est l'intérêt économique d'un tel processus ?

Construction d'un pentagone régulier inscrit dans un cercle

- UU' est un diamètre, OV est un rayon perpendiculaire
- ω est le milieu de OU'
- K est défini par $\omega K = \omega V$
- H est le milieu de OK
- HP est perpendiculaire à OU.



Alors l'angle $\widehat{\overrightarrow{OU},\overrightarrow{OP}}$ vaut $\frac{2\pi}{5}$, $OH = \cos\frac{2\pi}{5}$ et

UP est le côté du pentagone régulier.

Chapitre 3

Dénombrement, permutations, graphes

1. Ensembles finis

1.1 Définition et propriétés

Un ensemble E est *fini* s'il possède un nombre fini d'éléments.

Si E est un ensemble fini ayant au moins un élément, alors en numérotant ses éléments à partir de 1, on obtient $E=\{a_1,a_2,\ldots,a_n\}$. L'entier n est le nombre d'éléments de E.

Il y a en général plusieurs façons de numéroter les éléments de E, mais le nombre d'éléments ne dépend pas de la numérotation choisie. On convient que l'ensemble vide est fini et que son nombre d'éléments est zéro.

Notation : Notons $|E|$ le nombre d'éléments de l'ensemble fini E.

Des ensembles finis E et F ont le même nombre d'éléments si et seulement s'il existe une application bijective de E dans F.

Voici des propriétés très simples concernant les parties d'un ensemble fini.

Proposition. *Soit E un ensemble fini à n éléments.*

- *Toute partie de E est finie et possède au plus n éléments.*
- *Si A est une partie de E et si A possède n éléments, alors $A = E$.*

Proposition. *Si A et B sont des parties finies d'un ensemble, alors la partie $A \cup B$ est finie et l'on a*

$$|A \cup B| + |A \cap B| = |A| + |B|\,.$$

Démonstration. Si l'intersection $A\cap B$ est vide, le nombre d'éléments de $A\cup B$ est $|A|+|B|$ et la formule est vraie. Dans le cas général, les parties A et $B\setminus(A\cap B)$ sont disjointes et leur

réunion est $A \cup B$; on a donc $|A \cup B| = |A| + |B \setminus (A \cap B)|$. Les parties $A \cap B$ et $B \setminus (A \cap B)$ sont disjointes et leur réunion est B, donc on a $|B \setminus (A \cap B)| = |B| - |A \cap B|$, d'où la formule. ■

Proposition. *Soient E et F des ensembles finis.*

► *Le nombre de couples (x,y) tels que $x \in E$ et $y \in F$ est $|E||F|$: on a donc $|E \times F| = |E||F|$.*

► *Le nombre d'applications de E dans F est $|F|^{|E|}$.*

Démonstration. Posons $n = |E|$ et $p = |F|$. Il y a n façons de choisir un élément de l'ensemble E et pour chacun de ces choix, il y a p façons de choisir un élément de l'ensemble F. Il y a donc np couples (x,y) tels que $x \in E$ et $y \in F$. Pour définir une application $f : E \to F$, il faut, pour chaque élément $x \in E$, choisir son image $f(x)$ parmi les éléments de F : pour chaque élément de E, il y a donc p choix possibles. Puisque E possède n éléments, il y a $\underbrace{p \times p \times \cdots \times p}_{n \text{ facteurs}} = p^n$ choix pour définir une application de E dans F. ■

Dans un ensemble fini, il y a évidemment un nombre fini de parties : par exemple, l'ensemble $\{a,b,c\}$ a pour parties $\varnothing$, $\{a\}$, $\{b\}$, $\{c\}$, $\{a,b\}$, $\{a,c\}$, $\{b,c\}$ et $\{a,b,c\}$.

Corollaire. *Dans un ensemble à n éléments, il y a 2^n parties.*

Démonstration. Soit E un ensemble à n éléments et soit P l'ensemble des parties de E. Nous allons établir une bijection entre P et l'ensemble $\mathscr{F}$ des applications de E dans $\{0,1\}$. Puisqu'il y a $\big|\{0,1\}\big|^{|E|} = 2^n$ éléments dans l'ensemble $\mathscr{F}$, cela démontrera le corollaire.
À toute partie A de E, associons la fonction $c_A : E \to \{0,1\}$ définie en posant $c_A(x) = 1$ si $x \in A$, $c_A(x) = 0$ si $x \notin A$. La fonction c_A s'appelle la *fonction caractéristique de la partie A*.
À toute fonction $f : E \to \{0,1\}$, associons la partie X_f de E formée des éléments $x \in E$ tels que $f(x) = 1$.
La fonction caractéristique de la partie X_f est f et si A est une partie de E, alors la partie associée à la fonction c_A est A. L'application $A \mapsto c_A$ est donc une bijection de P dans $\mathscr{F}$ et l'application $f \mapsto X_f$ est la bijection réciproque. ■

1.2 Applications entre ensembles finis

Soient E et F des ensembles finis et $f : E \to F$ une application.
Rappelons que l'image de l'application f est l'ensemble, noté $f(E)$, des éléments $f(x)$, où x parcourt E. Les éléments $f(x)$ ne sont pas nécessairement tous différents (si f est une application constante, ils sont tous égaux), mais en tous cas, leur nombre est inférieur ou égal au nombre d'éléments de E. On a donc $\big|f(E)\big| \leqslant |E|$. De plus, $f(E)$ étant une partie de F, on a $\big|f(E)\big| \leqslant |F|$.
Résumons ces propriétés :

► *Le nombre d'éléments dans l'image de f est inférieur ou égal à $|E|$ et à $|F|$.*

► *On a $f(E) = F$ si et seulement si $\big|f(E)\big| = |F|$.*

En particulier, si E a strictement moins d'éléments que F, il existe au moins un élément de F qui n'a pas d'antécédent par f.

Proposition. *Soient E et $F = \{b_1 \ldots, b_n\}$ des ensembles finis et soit $f : E \to F$ une application. Pour chaque entier $i \in \{1, \ldots, n\}$, soit A_i l'ensemble des antécédents de b_i. Alors on a $|A_1| + \cdots + |A_n| = |E|$.*

Démonstration. On sait (page 21) que les parties A_i et A_j n'ont pas d'élément en commun si $i \neq j$ et que la réunion des parties A_i est l'ensemble E tout entier. Le nombre d'éléments de E est donc la somme $|A_1| + \cdots + |A_n|$. ■

Voici des propriétés importantes concernant les applications entre des ensembles finis ayant le même nombre d'éléments.

Proposition. *Soient E et F des ensembles finis ayant le même nombre d'éléments et soit $f : E \to F$ une application.*

i) *L'application f est une bijection si et seulement si $\big|f(E)\big| = |F|$.*

ii) *L'application f est une bijection si et seulement si tout élément de F a au plus un antécédent par f.*

Démonstration. Si f est une bijection, alors par définition, on a $f(E) = F$ et tout élément de F a exactement un antécédent par f.
Posons $F = \{b_1, \ldots, b_n\}$ et pour tout entier i compris entre 1 et n, notons A_i l'ensemble des antécédents de b_i. D'après la proposition précédente, on a $|E| = |A_1| + \cdots + |A_n|$ et comme $|E| = n$ par hypothèse, il vient

$$(*) \qquad \big(|A_1| - 1\big) + \cdots + \big(|A_n| - 1\big) = |A_1| + \cdots + |A_n| - n = |E| - n = 0$$

Supposons que $f(E)$ et F ont le même nombre d'éléments. On a donc $f(E) = F$, car $f(E)$ est une partie de F. Puisque tout élément de $f(E)$ possède par définition un antécédent par f, on en déduit que tout élément de F possède au moins un antécédent, autrement dit nous avons $|A_i| \geqslant 1$ pour tout i. Chacun des entiers $|A_i| - 1$ est positif ou nul, leur somme est nulle d'après $(*)$, donc on a $|A_i| - 1 = 0$ pour tout i : chaque partie A_i possède donc un et un seul élément. Cela veut dire que chaque élément de F a un et un seul antécédent, donc l'application f est bijective. Nous avons ainsi démontré la propriété (i).
Supposons maintenant que chaque partie A_i possède au plus un élément, donc $|A_i| \leqslant 1$ pour tout i. Les entiers $|A_i| - 1$ sont négatifs ou nuls, leur somme est nulle d'après $(*)$, donc ils sont tous nuls. Ainsi l'on a $|A_i| = 1$ pour tout i et l'on conclut comme précédemment que f est une bijection, ce qui démontre (ii). ■

Principe des tiroirs. *Soient E et F des ensembles finis et soit $f : E \to F$ une application. Si $|E| > |F|$, alors il existe des éléments a et a' appartenant à E tels que $a \neq a'$ et $f(a) = f(a')$.*

Démonstration. Reprenons les notations introduites dans la démonstration précédente et supposons $|E| > |F|$. On a donc $|A_1| + \cdots + |A_n| = |E| > n$. Puisque les nombres $|A_i|$ sont des entiers positifs ou nuls, on en déduit que l'un au moins est strictement supérieur à 1. Soit i tel que $|A_i| > 1$. Alors il existe dans A_i au moins deux éléments a et a' différents et l'on a $f(a) = b_i$, $f(a') = b_i$, donc $f(a) = f(a')$. ■

Cette propriété s'appelle le principe des tiroirs, car lorsqu'on range plus de n objets dans n tiroirs, l'un des tiroirs doit contenir au moins deux objets.

1.3 Des dénombrements utiles

Si des ensembles finis E et F ont le même nombre d'éléments, on sait qu'il existe des applications bijectives de E dans F. Calculons le nombre de ces bijections.

Proposition. *Si E et F sont des ensembles finis ayant le même nombre n d'éléments, il y a $n!$ applications bijectives de E dans F.*

Démonstration. Posons $E = \{a_1, \dots, a_n\}$ et $F = \{b_1, \dots, b_n\}$. Pour définir une bijection de $f : E \to F$, on peut choisir l'élément $f(a_1)$ arbitrairement dans F ; l'élément $f(a_2)$ doit être différent de $f(a_1)$, ce qui laisse $n-1$ possibilités ; ensuite, il reste $n-2$ possibilités pour choisir $f(a_2)$, et en général $n-k$ possibilités pour choisir $f(a_k)$. Finalement, il y a $n(n-1)\cdots 2 \cdot 1$ possibilités pour définir une bijection de E dans F. ■

Quand on se donne un ensemble fini E sous la forme $E=\{a_1,\dots,a_n\}$, les éléments de E ont été ordonnés : il y a un premier élément a_1, un deuxième a_2, etc. Si l'on change la numérotation, les éléments de E sont les mêmes, mais l'ordre est différent. Numéroter les éléments de E, c'est établir une bijection de l'ensemble $\{1,2,\dots,n\}$ dans E. D'après la proposition précédente, il y a donc $n!$ façons de numéroter les éléments de E.

Nombre de p-arrangements

Définition

Soit E un ensemble fini à n éléments et soit p un entier tel que $1 \leqslant p \leqslant n$. Un *$p$-arrangement d'éléments de E* est une suite $(a_1, a_2, \dots, a_p)$ de p éléments de E deux à deux différents.

Pour définir un p-arrangement $(a_1, a_2, \dots, a_p)$ d'un ensemble E à n éléments, il y a n façons de choisir a_1, $n-1$ façons de choisir a_2 (car a_2 doit être différent de a_1), etc, donc finalement $n(n-1)\cdots(n-p+1)$ façons de choisir les éléments $a_1, \dots, a_p$.

Proposition. *Soit E un ensemble à n éléments. Si p est un entier tel que $1 \leqslant p \leqslant n$, le nombre de p-arrangements de E est $n(n-1)\cdots(n-p+1) = \dfrac{n!}{(n-p)!}$.*

Si E est un ensemble à n éléments, un n-arrangement de E est une bijection de $\{1, 2, \dots, n\}$ dans E : on retrouve ainsi qu'il y a $n!$ bijections entre deux ensembles à n éléments. Ainsi, on a l'égalité $0! = 1$.

Nombre de parties à p éléments

Proposition. *Soit E un ensemble à n éléments. Si p est un entier tel que $0 \leqslant p \leqslant n$, le nombre de parties de E ayant p éléments est $\dfrac{n!}{p!(n-p)!}$.*

Démonstration. Il n'y a qu'une partie à zéro éléments : la partie vide ; le résultat est donc vrai si $p = 0$, car par convention, $0! = 1$. Supposons $p \geqslant 1$. Soit A une partie de E ayant p éléments. Les p-arrangements de E formés avec les éléments de A sont définis en se donnant une bijection de $\{1, 2, \dots, p\}$ dans A. Il y a $p!$ bijections de $\{1, 2, \dots, p\}$ dans A, donc il y a

$p!$ p-arrangements formés avec les éléments de A. Comme le nombre de p-arrangements de E est $\frac{n!}{(n-p)!}$, on en déduit qu'il y a $\frac{1}{p!}\frac{n!}{(n-p)!}$ parties de E à p éléments. ■

Définition
Soit n un entier positif ou nul et soit p un entier relatif. Si $0\leqslant p\leqslant n$, l'entier $\frac{n!}{p!(n-p)!}$ s'appelle un *nombre binomial* et se note $\binom{n}{p}$. Si $p<0$ ou si $p>n$, on pose $\binom{n}{p}=0$.

- On a $\binom{n}{0}=\binom{n}{n}=1$ et $\binom{n}{1}=\binom{n}{n-1}=n$, car dans un ensemble E à n éléments, l'unique partie à 0 éléments est l'ensemble vide, l'unique partie à n éléments est E lui-même et il y a n parties à 1 élément.
- On a $\binom{n}{p}=\binom{n}{n-p}$: en effet, si à toute partie A de E on associe son complémentaire $E\setminus A$, on réalise une bijection entre les parties à p éléments et les parties à $n-p$ éléments.

Nous avons montré dans le premier paragraphe qu'un ensemble à n éléments possède 2^n parties. On a donc l'égalité

$$\binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{n-1}+\binom{n}{n}=2^n$$

Voici une importante propriété des coefficients binomiaux ; elle se vérifie par simple calcul à partir de la définition.

Proposition. *Pour tous entiers $n\geqslant 1$ et $p\in\mathbb{Z}$, on a $\binom{n}{p}=\binom{n-1}{p}+\binom{n-1}{p-1}$.*

Formule du binôme de Newton. *Soient a et b des nombres réels ou complexes, ou bien des polynômes à coefficients réels ou complexes. Pour tout entier $n\geqslant 2$, on a l'égalité*

$$(a+b)^n=a^n+\binom{n}{1}a^{n-1}b+\binom{n}{2}a^{n-2}b^2+\cdots+\binom{n}{k}a^{n-k}b^k+\cdots+\binom{n}{n-1}ab^{n-1}+b^n.$$

- Pour $n=2$, on retrouve la formule connue $(a+b)^2=a^2+\binom{2}{1}ab+b^2=a^2+2ab+b^2$.
- On a $\binom{3}{1}=\binom{3}{2}=3$, d'où $(a+b)^3=a^3+3a^2b+3ab^2+b^3$, formule qu'il est bon de connaître.
- On a de même $\binom{4}{1}=\binom{4}{3}=4$ et $\binom{4}{2}=\frac{4\times 3}{2}=6$, donc $(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4$.

Démonstration. Développons le produit $(a+b)^n=\underbrace{(a+b)(a+b)\cdots(a+b)}_{n \text{ facteurs}}$ en une somme de termes. Un terme quelconque du développement s'obtient en choisissant, dans chaque facteur $(a+b)$, l'un des nombres a ou b et en faisant le produit : si l'on choisit le terme a dans k des facteurs, on obtient le terme a^kb^{n-k} dans le développement. Pour tout entier k compris entre 0 et n, il y a $\binom{n}{k}$ façons de choisir k facteurs parmi n, donc $\binom{n}{k}$ termes a^kb^{n-k} dans le développement de $(a+b)^n$. Le produit $(a+b)^n$ est donc égal à la somme des $\binom{n}{k}a^kb^{n-k}$, où k parcourt les entiers compris entre 0 et n. Si $k=0$, on a $\binom{n}{k}a^kb^{n-k}=\binom{n}{0}a^0b^n=b^n$,

car $\binom{n}{0} = 1$ et $a^0 = 1$; si $k = n$, alors $\binom{n}{k}a^k b^{n-k} = \binom{n}{n}a^n b^0 = a^n$, car $\binom{n}{n} = 1$; dans le cas général, on a $\binom{n}{k}a^k b^{n-k} = \binom{n}{n-k}a^k b^{n-k}$, d'où la formule. ■

Corollaire. *Soient n et p des entiers tels que $1 \leqslant p \leqslant n$. Il y a $\binom{n}{p}$ applications strictement croissantes de $\{1, \dots, p\}$ dans $\{1, \dots, n\}$.*

Démonstration. Si u est une application strictement croissante de $\{1,\dots,p\}$ dans $\{1,\dots,n\}$, on a $u(1) < u(2) < \cdots < u(p)$, donc les éléments $u(1), \dots, u(p)$ sont deux à deux différents et l'ensemble $\{u(1), \dots, u(p)\}$ est une partie à p éléments de $\{1, \dots, n\}$.
Notons S l'ensemble des applications strictement croissantes de $\{1, \dots, p\}$ dans $\{1, \dots, n\}$ et P l'ensemble des parties à p éléments de $\{1, \dots, n\}$. À toute application strictement croissante $u : \{1, \dots, p\} \to \{1, \dots, n\}$, associons la partie $\{u(1), \dots, u(p)\}$. On définit ainsi une application $f : S \to P$. Montrons que f est une bijection, ce qui prouvera le corollaire.
Soit U une partie à p éléments de $\{1, \dots, n\}$. Rangeons les éléments de U dans l'ordre croissant : on obtient $U = \{u_1, \dots, u_p\}$, où les entiers u_i vérifient $1 \leqslant u_1 < \cdots < u_p \leqslant n$. Définissons l'application $u : \{1, \dots, p\} \to \{1, \dots, n\}$ en posant $u(i) = u_i$ si $1 \leqslant i \leqslant p$. L'application u est strictement croissante et $\{u(1), \dots, u(p)\} = U$, autrement dit u est un antécédent de U par f.
Supposons que v est un (autre) antécédent de U. On a $v \in S$, donc $v : \{1, \dots, p\} \to \{1, \dots, n\}$ est une application strictement croissante ; de plus, on a $f(v) = U$, donc $\{v(1), \dots, v(p)\} = U = \{u(1), \dots, u(p)\}$. Puisque v est strictement croissante, on a $v(1) < \cdots < v(p)$. On en déduit les égalités $v(1) = u(1), \dots, v(p) = u(p)$, donc $v = u$.
Par l'application $f : S \to P$, tout élément de P possède exactement un antécédent. L'application f est donc bijective. ■

Nombre d'applications croissantes

Dans ce paragraphe, p est un entier au moins égal à 1 et n est un entier positif ou nul. Considérons les inéquations suivantes

$$(1) \qquad x_1 + x_2 + \cdots + x_p \leqslant n \quad , \qquad x_i \in \mathbb{N} \ , \ x_i \geqslant 1$$
$$(2) \qquad x_1 + x_2 + \cdots + x_p \leqslant n \quad , \qquad x_i \in \mathbb{N}$$

où l'inconnue est une suite $(x_1, \dots, x_p)$ d'entiers positifs ou nuls pour (2), d'entiers strictement positifs pour (1).
Voici deux interprétations utiles pour les solutions de ces inéquations.

A) *Il y a une bijection entre les solutions de (2) et les applications croissantes de $\{1, \dots, p\}$ dans $\{0, 1, \dots, n\}$.*

Si $s = (x_1, \dots, x_p)$ est une solution de (2), on définit l'application

$$u_s : \{1, \dots, p\} \to \{0, 1, \dots, n\}$$

en posant $u_s(k) = x_1 + \cdots + x_k$ si $1 \leqslant k \leqslant p$. Puisque les entiers x_k sont positifs ou nuls, l'application u_s est croissante.
Réciproquement, supposons que v est une application croissante de $\{1, \dots, p\}$ dans $\{0, 1, \dots, n\}$; posons $x_1 = v(1)$ et $x_k = v(k) - v(k-1)$ pour $2 \leqslant k \leqslant p$; par hypothèse, on a $x_1 \geqslant 0$ et $x_k = v(k) - v(k-1) \geqslant 0$ pour $2 \leqslant k \leqslant p$, donc $(x_1, \dots, x_p)$ est une

suite d'entiers positifs ou nuls ; pour tout entier k tel que $1 \leqslant k \leqslant p$, on a

$$x_1 + \cdots + x_k = v(1) + \big(v(2) - v(1)\big) + \cdots + \big(v(k) - v(k-1)\big) = v(k) \leqslant n\,.$$

La suite $s = (x_1, \ldots, x_p)$ est donc une solution de (2) et l'on a $v = u_s$.

B) *Il y a une bijection entre les solutions de l'inéquation* (1) *et les applications strictement croissantes* $\{1, \ldots, p\} \to \{1, \ldots, n\}$.

Remarquons que toute solution de (1) est aussi une solution de (2). Supposons que $s=(x_1,\ldots,x_p)$ est une solution de (1). Puisque les entiers x_k sont tous au moins égaux à 1, on a $n \geqslant x_1 + \cdots + x_p \geqslant p$, donc $n \geqslant p \geqslant 1$. L'application u_s est strictement croissante et comme on a $u_s(1) = x_1 \geqslant 1$, u_s prend ses valeurs dans l'ensemble $\{1, \ldots, n\}$. Réciproquement, si u_s prend ses valeurs dans $\{1, \ldots, n\}$ et est strictement croissante, alors on a $x_1 = u_s(1) \geqslant 1$ et $x_k = u_s(k) - u_s(k-1) \geqslant 1$ pour tout entier k compris entre 2 et p.

Proposition

i) *L'inéquation* (1) *possède* $\binom{n}{p}$ *solutions.*

ii) *L'inéquation* (2) *possède* $\binom{p+n}{p}$ *solutions.*

iii) *Si* $n \geqslant 1$, *le nombre d'applications croissantes de* $\{1, \ldots, p\}$ *dans* $\{1, \ldots, n\}$ *est* $\binom{p+n-1}{p}$.

Démonstration. On sait qu'il y a $\binom{n}{p}$ applications strictement croissantes de $\{1,\ldots,p\}$ dans $\{1, \ldots, n\}$, donc l'inéquation (1) possède $\binom{n}{p}$ solutions.
On peut facilement passer d'une solution de (1) à une solution de (2) et vice-versa : si $(x_1, \ldots, x_p)$ est une suite d'entiers, posons $y_i = 1 + x_i$ pour $1 \leqslant i \leqslant p$. On a alors $y_1 + \cdots + y_p = p + (x_1 + \cdots + x_p)$, d'où l'équivalence

$$x_1 + \cdots + x_p \leqslant n \ \text{ et } \ x_i \geqslant 0 \text{ pour tout } i$$
$$\iff \ y_1 + \cdots + y_p \leqslant p + n \ \text{ et } \ y_i \geqslant 1 \text{ pour tout } i.$$

Les solutions de l'inéquation (2) sont donc en bijection avec les solutions de l'inéquation $y_1 + \cdots + y_p \leqslant p + n$ telles que $y_i \geqslant 1$ pour tout i. C'est une inéquation du type (1) (où n est remplacé par $p + n$) : d'après (i), il y a donc $\binom{p+n}{p}$ solutions à l'inéquation (2).
Ainsi le nombre d'applications croissantes de $\{1, \ldots, p\}$ dans $\{0, \ldots, n\}$ est aussi $\binom{p+n}{p}$. Puisque $\{0, \ldots, n\}$ possède $n+1$ éléments, le nombre d'applications croissantes de $\{1, \ldots, p\}$ dans $\{1, \ldots, n\}$ est $\binom{p+(n-1)}{p}$. ■

Exemples

- Les triplets (x, y, z) d'entiers tels que $x \geqslant 1$, $y \geqslant 1$, $z \geqslant 1$ et $x + y + z \leqslant 5$ sont :

  $(1, 1, 3), (1, 3, 1), (3, 1, 1), (2, 2, 1), (2, 1, 2), (1, 2, 2)$ pour lesquels $x + y + z = 5$,

  $(1, 1, 2), (1, 2, 1), (2, 1, 1)$ pour lesquels $x + y + z = 4$,

  et $(1, 1, 1)$ pour lequel $x + y + z = 3$.

  Il y en a au total $\binom{5}{3} = \dfrac{5 \times 4}{2} = 10$.
- Les applications croissantes de $\{1,2,3\}$ dans $\{1,2\}$ sont : l'application c_1 constante de valeur 1, définie par $c_1(1) = c_1(2) = c_1(3) = 1$; l'application c_2 constante de

valeur 2 ; l'application c_3 définie par $c_3(1)=c_3(2)=1$ et $c_3(3)=2$; et l'application c_4 définie par $c_4(1)=1$ et $c_4(2)=c_4(3)=2$. Leur nombre est bien $\binom{3+2-1}{3}=\binom{4}{3}=4$.

Corollaire. *Soient n et p des entiers au moins égaux à 1.*

i) *L'équation $x_1+\cdots+x_p=n$, où $x_i \in \mathbb{N}$ et $x_i \geqslant 1$, possède $\binom{n-1}{p-1}$ solutions.*

ii) *L'équation $x_1+\cdots+x_p=n$, où $x_i \in \mathbb{N}$, possède $\binom{p+n-1}{p-1}$ solutions.*

Démonstration. Si $x_1,\ldots,x_p$ sont des entiers, on a l'équivalence

$$x_1+\cdots+x_p=n \iff \big(x_1+\cdots+x_p \leqslant n \text{ et } x_1+\cdots+x_p > n-1\big).$$

Le nombre de solutions de l'équation $x_1+\cdots+x_p=n$, où $x_i \in \mathbb{N}$ et $x_i \geqslant 1$, est donc $\binom{n}{p}-\binom{n-1}{p}=\binom{n-1}{p-1}$. De même, le nombre de solutions de l'équation $x_1+\cdots+x_p=n$, où $x_i \in \mathbb{N}$, est $\binom{p+n}{p}-\binom{p+n-1}{p}=\binom{p+n-1}{p-1}$. ∎

Exemple. Pour organiser un jeu publicitaire dans un centre commercial, on prépare p lots différents. En plus du lot, chaque gagnant se verra offrir au moins deux bons d'achat. On dispose de n bons d'achat, tous du même montant. Combien y a-t-il de façons de répartir tous les bons entre les lots ?
Numérotons les lots de 1 à p et notons x_i le nombre de bons d'achat donnés avec le lot numéro i. Les bons étant tous les mêmes, une répartition est déterminée par les entiers $x_1,x_2,\ldots,x_p$. Puisque tous les bons sont utilisés, la somme des x_i est égale à n. Le nombre cherché est donc le nombre de solutions $(x_1,\ldots,x_p)$ de l'équation $x_1+\cdots+x_p=n$ telles que $x_i \geqslant 2$ pour tout i.
En posant $y_i=x_i-2$, l'équation $x_1+\cdots+x_p=n$ est équivalente à $y_1+\cdots+y_p=n-2p$, où les y_i sont des entiers positifs ou nuls. Si $n<2p$, il n'y a pas de solution. Si $n \geqslant 2p$, le nombre de solutions est $\binom{p+(n-2p)-1}{p-1}=\binom{n-p-1}{p-1}$.

Applications

Nombre de rangements de n objets dans p boîtes

Numérotons les boîtes de 1 à p et notons x_i le nombre d'objets dans la i-ième boîte. En ne tenant compte que du nombre d'objets dans chaque boîte, les solutions sont les suites $(x_1,\ldots,x_p)$ d'entiers positifs ou nuls tels que $x_1+\cdots+x_p=n$. Le nombre de possibilités est donc $\binom{p+n-1}{p-1}$.

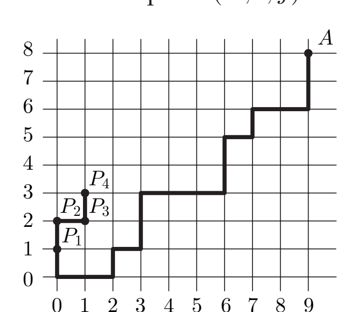
Combinaisons avec répétitions

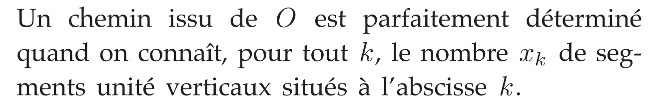
Étant donné un ensemble E à n éléments, une *p-combinaison avec répétitions* de ces n éléments est un ensemble à p éléments formé de clones d'éléments de E.
Si par exemple $E=\{a,b,c,d,e,f\}$, alors a,a,c,d,d,d,f représente une 7-combinaison avec répétitions des 6 éléments de E. De même, u,u,u,v,v est une 5-combinaison avec répétitions des deux éléments u et v.
Une p combinaison avec répétitions des n éléments $a_1,\ldots,a_n$ est déterminée par le nombre de fois qu'y figure chacun des élément $a_1,\ldots,a_n$: si a_1 est répété x_1 fois, si a_2 est répété x_2 fois et a_k, x_k fois, alors $x_1+x_2+\cdots+x_n=p$. Il y a donc autant de p combinaisons avec répétitions de n éléments que de solutions

de l'équation $x_1 + x_2 + \cdots + x_n = p$, où les x_i sont entiers positifs ou nuls. Le nombre de p combinaisons avec répétitions de n éléments est $\binom{n+p-1}{n-1}$.
C'est aussi le nombre de rangements de n objets dans p boîtes.

Nombre de chemins sur un quadrillage

Repérons les points du plan par leurs coordonnées dans un repère $(O; \vec{i}, \vec{j})$ et considérons le quadrillage formé des droites parallèles aux axes. Appelons *chemin issu de* O toute suite $P_0 = O, P_1, \ldots, P_m$ de points tels que, pour tout k, le vecteur $\overrightarrow{P_k P_{k+1}}$ est égal à $\vec{i}$ ou à $\vec{j}$. Les coordonnées des points P_k sont donc des entiers positifs ou nuls. La suite $O, P_1, \ldots, P_m$ définit une ligne brisée portée par le quadrillage.





Un chemin issu de O est parfaitement déterminé quand on connaît, pour tout k, le nombre x_k de segments unité verticaux situés à l'abscisse k.
Par exemple, le chemin de O à A représenté sur la figure correspond à la suite

$$x_0 = x_1 = 0, x_2 = 1, x_3 = 2, x_4 = x_5 = 0, x_6 = 2, x_7 = 1, x_8 = 0, x_9 = 2\,.$$

Soit A le point de coordonnées (p, q), où p et q sont des entiers positifs ou nuls. Pour qu'un chemin issu de O ait le point A comme extrémité, il faut et il suffit que l'on ait $x_0 + x_1 + \cdots + x_p = q$. Les chemins issus de O et d'extrémité A sont donc en bijection avec les solutions de l'équation $x_0 + \cdots + x_p = q$, où $x_k \in \mathbb{N}$. D'après le corollaire précédent, il y a $\binom{(p+1)+q-1}{(p+1)-1} = \binom{p+q}{p}$ chemins issus de O et d'extrémité A.

Applications ayant pour image leur ensemble d'arrivée

Soient E et F des ensembles finis. Pour qu'il existe une application $f : E \to F$ telle que $f(E) = F$, il faut que le nombre d'éléments de E soit supérieur ou égal au nombre d'éléments de F (voir page 58).
Réciproquement, si l'on a $|E| \geqslant |F|$, il est facile de construire une application $f : E \to F$ telle que $f(E) = F$: par exemple, si $E = \{a_1, \ldots, a_p\}$, $F = \{b_1, \ldots, b_n\}$ et $p \geqslant n$, on peut poser $f(a_i) = b_i$ si $1 \leqslant i \leqslant n$ et $f(a_i) = b_n$ si $n < i \leqslant p$.
Supposons désormais $|E| \geqslant |F|$ et comptons le nombre d'applications de E dans F dont l'image $f(E)$ est l'ensemble d'arrivée F tout entier :

> *une telle application permet de parcourir <u>tous</u> les éléments $y = f(x)$ de l'ensemble d'arrivée en faisant varier x dans l'ensemble de départ.*

On dit que c'est une application *surjective*.

Préliminaire. *Si A est un ensemble fini non vide, il y a autant de parties de A ayant un nombre pair d'éléments que de parties de A ayant un nombre impair d'éléments.*

Démonstration. L'ensemble A étant par hypothèse non vide, on peut choisir un élément $a \in A$. Posons $B = A \setminus \{a\}$. Les parties de A qui ne contiennent pas a sont les parties de

B. Les parties de A contenant a sont de la forme $Y = X \cup \{a\}$, où X est une partie de B ; puisque $a \notin X$, le nombre d'éléments de Y est $1 + |X|$.

Notons p_B le nombre de parties de B ayant un nombre pair d'éléments et i_B le nombre de parties de B ayant un nombre impair d'éléments. Une partie de A ayant un nombre pair d'éléments est ou bien une partie de B, ou bien de la forme $X \cup \{a\}$, où X est une partie de B ayant un nombre impair d'éléments. On en déduit qu'il y a $p_B + i_B$ parties de A ayant un nombre pair d'éléments. De même, il y a $i_B + p_B$ parties de A ayant un nombre impair d'éléments. Le résultat s'ensuit. ■

Si A est l'ensemble vide, la seule partie de A est A lui-même : il n'y a pas de parties de A ayant un nombre impair d'éléments et il y a une seule partie dont le nombre d'éléments est pair (en fait égal à 0).

Proposition. *Soient E un ensemble à p éléments et F un ensemble à n éléments, où $p \geqslant n \geqslant 0$. Le nombre d'applications $f : E \to F$ telles que $f(E) = F$ est*

$$n^p - \binom{n}{1}(n-1)^p + \binom{n}{2}(n-2)^p + \cdots + (-1)^k\binom{n}{k}(n-k)^p + \cdots + (-1)^{n-1}\binom{n}{n-1}.$$

On peut aussi calculer ces nombres de proche en proche : voir l'exercice 8 en fin de chapitre.

Démonstration. Pour tout ensemble A, posons $d_A = \sum_{X \subset A} (-1)^{|X|}$, le signe $\sum_{X \subset A}$ voulant dire que l'on effectue la sommation sur toutes les parties X de l'ensemble A. Si X est une partie de A, alors $(-1)^{|X|} = 1$ si le nombre d'éléments de X est pair et $(-1)^{|X|} = -1$ si le nombre d'éléments de X est impair. D'après le résultat préliminaire, on a donc $d_A = 0$ si A est non vide. Si A est l'ensemble vide, alors $d_A = (-1)^0 = 1$. Pour une application $f : E \to F$, on a donc $d_{F \setminus f(E)} = 1$ si $f(E) = F$ et $d_{F \setminus f(E)} = 0$ sinon. Notons S l'ensemble des applications $f : E \to F$ telles que $f(E) = F$. En notant F^E l'ensemble des applications de E dans F, nous venons de montrer que le nombre d'éléments de S est

$$|S| = \sum_{f \in F^E} d_{F \setminus f(E)} = \sum_{f \in F^E} \sum_{X \subset F \setminus f(E)} (-1)^{|X|}$$

Si $f : E \to F$ est une application et si X est une partie de F, on a l'équivalence

$$X \subset F \setminus f(E) \iff f(E) \subset F \setminus X\,.$$

La seconde propriété signifie que f est une application de E dans $F \setminus X$. Il vient donc

$$|S| = \sum_{X \subset F} \sum_{f \in (F \setminus X)^E} (-1)^{|X|} = \sum_{X \subset F} (-1)^{|X|} |F \setminus X|^p$$

car il y a $|F \setminus X|^p$ applications de E dans $F \setminus X$. Pour tout entier k tel que $0 \leqslant k \leqslant n$, posons

$$s_k = \sum_{\substack{X \subset F \\ |X| = k}} (-1)^k |F \setminus X|^p = (-1)^k \sum_{\substack{X \subset F \\ |X| = k}} |F \setminus X|^p\,,$$

de sorte que l'on a $|S| = \sum_{0 \leqslant k \leqslant n} s_k$. Si X est une partie de F à k éléments, $F \setminus X$ possède $n-k$ éléments, et comme il y a $\binom{n}{k}$ parties de F à k éléments, il vient $s_k = (-1)^k \binom{n}{k}(n-k)^p$. On en déduit $S = \sum_{0 \leqslant k \leqslant n} (-1)^k \binom{n}{k}(n-k)^p$ qui est la formule annoncée. ■

Exemple

- Sur les $3^4=81$ applications de $\{1,2,3,4\}$ dans $\{1,2,3\}$, il y en a $3^4-3\times 2^4+3=81-3\times 16+3=36$ dont l'image est l'ensemble $\{1,2,3\}$.
- Si $n=p$, les applications f de E dans F telles que $f(E)=F$ sont les bijections : on sait qu'il y en a $n!$.

1.4 Probabilité binomiale et loi des grands nombres

Supposons qu'au cours d'épreuves indépendantes, un événement se produit avec la probabilité p (p est un nombre réel tel que $0\leqslant p\leqslant 1$).
Numérotons les épreuves de 1 à n et donnons-nous une partie A à k éléments de l'ensemble $\{1,2,\dots,n\}$. Pour que l'événement se produise exactement quand le numéro de l'épreuve est dans A, la probabilité est $p^k(1-p)^{n-k}$. Comme il y a $\binom{n}{k}$ parties à k éléments, on en déduit que *la probabilité pour que l'événement se produise exactement k fois en n épreuves est*

$$P(k)=\binom{n}{k}p^k(1-p)^{n-k}.$$

Pour n et p fixés, les points de coordonnées $\big(k,P(k)\big)$ se trouvent sur une courbe en forme de cloche.

$\boldsymbol{n=10, p=0{,}4}$ $\boldsymbol{n=10, p=0{,}8}$

Par définition des probabilités $P(k)$, on a bien sûr $P(0)+P(1)+\dots+P(n)=1$, égalité qui n'est autre que l'identité $\big[p+(1-p)\big]^n=1$.
Notons X la variable aléatoire : nombre de fois que l'événement se produit en n épreuves.

L'espérance. *L'espérance de X* est le nombre $E=0P(0)+1P(1)+2P(2)+\dots+nP(n)$. Intuitivement, l'espérance est le nombre de succès auquel on peut s'attendre au cours des n épreuves.

Puisqu'on a $k\binom{n}{k}=n\binom{n-1}{k-1}$, il vient $E=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}=np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}$.

D'après la formule du binôme, on a $\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}=\big[1+(1-p)\big]^{n-1}=1$, d'où

$$E=np$$

C'est pour l'entier k le plus proche de E que $P(k)$ est maximum (voir les courbes ci-dessus).

La variance. *La variance de X* est le nombre $V=\sum_{k=0}^{n}(k-E)^2P(k)$, autrement dit l'espérance de la variable aléatoire $(X-E)^2$ (carré de l'écart à l'espérance). Comme indice de dispersion des résultats, on choisit souvent *l'écart-type* $\sigma=\sqrt{V}$. Montrons

que l'on a
$$V = np(1-p)$$

En utilisant l'identité $k^2 = k(k-1)+k$, il vient $(k-np)^2 = k(k-1)+(1-2np)k+n^2p^2$ et en posant $s = \sum_{k=2}^{n} k(k-1)P(k)$, on obtient $V = s+(1-2np)E+n^2p^2 = s+np-n^2p^2$. Puisque $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$, on a $s = n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{n-k}$; comme la somme vaut $\big[p+(1-p)\big]^{n-2} = 1$ d'après la formule du binôme, on trouve finalement $V = n(n-1)p^2 + np - n^2p^2 = -np^2 + np = np(1-p)$.

Exemple. La désintégration d'une substance radioactive est régie par la loi suivante : s'il y a N atomes au départ, la probabilité pour que n d'entre eux se désintègrent pendant l'intervalle de temps $[0,t]$ est $\binom{N}{n}(1-e^{-\lambda t})^n e^{-\lambda t(N-n)}$, où la constante de désintégration λ est un nombre positif caractéristique de la substance. Il s'agit donc d'une loi binomiale de probabilité $e^{-\lambda t}$.
Supposons qu'on a la probabilité c de détecter une désintégration (c dépend de la position du compteur par rapport au matériel radioactif). Sur n désintégrations produites, la probabilité d'en détecter k est $\binom{n}{k}c^k(1-c)^{n-k}$.
Soit X la variable aléatoire : nombre de désintégrations observées pendant l'intervalle de temps $[0,t]$. La probabilité pour que ce nombre soit k est
$$P(X=k) = \sum_{n=k}^{N} \binom{N}{n}(1-e^{-\lambda t})^n e^{-\lambda t(N-n)} \binom{n}{k}c^k(1-c)^{n-k}$$
Comme on a $\binom{N}{n}\binom{n}{k} = \binom{N}{k}\binom{N-k}{n-k}$, il vient
$$\begin{aligned} P(X=k) &= \tbinom{N}{k}c^k(1-e^{-\lambda t})^k \textstyle\sum_{n=k}^{N} \tbinom{N-k}{n-k}(1-e^{-\lambda t})^{n-k}(1-c)^{n-k}e^{\lambda t(n-k)}e^{-\lambda t(N-k)} \\ &= \tbinom{N}{k}c^k(1-e^{-\lambda t})^k e^{-\lambda t(N-k)} \textstyle\sum_{i=0}^{N-k} \tbinom{N-k}{i}(1-e^{-\lambda t})^i(1-c)^i(e^{\lambda t})^i \end{aligned}$$
(on a sorti de la somme des facteurs qui ne dépendent pas de n et l'on a fait le changement d'indice $i = n-k$). En posant $a = (1-e^{-\lambda t})(1-c)e^{\lambda t} = -1+e^{\lambda t}+c(1-e^{\lambda t})$, la somme s'écrit (formule du binôme)
$$\sum_{i=0}^{N-k} \binom{N-k}{i}a^i = (1+a)^{N-k} = \big[e^{\lambda t}+c(1-e^{\lambda t})\big]^{N-k}.$$
On a ainsi
$$\begin{aligned} P(X=k) &= \tbinom{N}{k}c^k(1-e^{-\lambda t})^k(e^{-\lambda t})^{N-k}\big[e^{\lambda t}+c(1-e^{\lambda t})\big]^{N-k} \\ &= \tbinom{N}{k}c^k(1-e^{-\lambda t})^k\big[1-c(1-e^{-\lambda t})\big]^{N-k} = \tbinom{N}{k}r^k(1-r)^{N-k}, \end{aligned}$$
où $r = c(1-e^{-\lambda t})$. La variable aléatoire X suit donc une loi binomiale de probabilité r : pendant l'intervalle de temps $[0,t]$, on peut espérer détecter $E(t) = Nc(1-e^{-\lambda t})$ désintégrations ; en mesurant un rapport $E(t_2)/E(t_1)$, il est possible de calculer la constante de désintégration λ caractéristique de la substance radioactive.

Si t est assez petit, alors $1 - e^{-\lambda t} \sim \lambda t$ donc $r \sim c\lambda t$; dans ce cas, tout se passe comme si la constante de désintégration était $c\lambda$ et qu'on pouvait observer la totalité des désintégrations.

La loi des grands nombres

Répétons un grand nombre d'épreuves indépendantes pour un événement se produisant avec la probabilité p. Pour un nombre d'épreuves n fixé, la fréquence d'apparition de l'événement est une variable aléatoire : si l'événement s'est produit k fois, sa fréquence est $f_n = k/n$.
Si ε est un nombre entre 0 et 1, on note $P\big[|f_n - p| < \varepsilon\big]$ la probabilité pour que f_n soit comprise entre $p - \varepsilon$ et $p + \varepsilon$.
La loi des grands nombres affirme que

$$P\big[|f_n - p| < \varepsilon\big] \text{ tend vers } 1 \text{ quand } n \text{ tend vers l'infini.}$$

Pour rendre plus précise cette affirmation, donnons-nous un nombre δ entre 0 et 1 et cherchons à partir de quelle valeur de n on aura l'inégalité $P\big[|f_n - p| < \varepsilon\big] > 1 - \delta$.

Par définition, la probabilité $P\big[|f_n - p| < \varepsilon\big]$ est la somme $a_n = \sum_{|k/n-p|<\varepsilon} \binom{n}{k} p^k (1-p)^{n-k}$.
En posant $b_n = \sum_{|k/n-p|\geqslant\varepsilon} \binom{n}{k} p^k (1-p)^{n-k}$, il vient $a_n + b_n = 1$. D'après la définition de la variance, on a aussi $np(1-p) = V \geqslant \sum_{|k/n-p|\geqslant\varepsilon} |k - np|^2 \binom{n}{k} p^k (1-p)^{n-k}$; dans la somme, chaque facteur $|k - np| = n|k/n - p|$ est supérieur ou égal à $n\varepsilon$, donc

$$np(1-p) \geqslant n^2\varepsilon^2 \sum_{|k/n-p|\geqslant\varepsilon} \binom{n}{k} p^k (1-p)^{n-k} = n^2\varepsilon^2 b_n = n^2\varepsilon^2(1 - a_n)\,.$$

On en déduit $a_n \geqslant 1 - \dfrac{p(1-p)}{n\varepsilon^2}$. Pour tout entier n supérieur à $\dfrac{p(1-p)}{\delta\varepsilon^2}$, on a $\dfrac{p(1-p)}{n\varepsilon^2} < \delta$, donc $a_n > 1 - \delta$.
Puisqu'on a $0 < p < 1$, le produit $p(1-p)$ ne peut dépasser $1/4$ qui est le maximum de la fonction $x(1-x)$ quand x parcourt l'intervalle $[0,1]$. Si l'on prend $n > \dfrac{1}{4\delta\varepsilon^2}$, on aura aussi $n > \dfrac{p(1-p)}{\delta\varepsilon^2}$ et par conséquent $P\big[|f_n - p| < \varepsilon\big] > 1 - \delta$.

1.5 Espérance et variance d'une variable aléatoire discrète

Rappelons la définition générale de l'espérance et de la variance d'une variable aléatoire définie sur un ensemble fini d'événements.

Définitions

Soit Y une variable aléatoire prenant les valeurs $y_1, y_2, \ldots, y_n$. Notons p_i la probabilité pour que Y prenne la valeur y_i, probabilité qu'on note aussi $P(Y = y_i)$.
► *L'espérance* de Y est le nombre $E(Y) = \sum_{i=1}^{n} y_i\, p_i$.

► La *variance* de Y est l'espérance de la variable aléatoire $\big[Y - E(Y)\big]^2$, c'est-à-dire le nombre $V(Y) = \sum_{i=1}^n (y_i - m)^2 p_i$, où $m = E(Y)$. Le nombre $\sqrt{V(Y)}$ est *l'écart-type*.

Propriétés de l'espérance et de la variance. *Soient Y et Z des variables aléatoires prenant au plus un nombre fini de valeurs et soient a, b des nombres.*

i) $E(aY + bZ) = aE(Y) + bE(Z)$

ii) $V(Y) = E(Y^2) - \big[E(Y)\big]^2$

iii) $V(aY + b) = a^2 V(Y)$.

Démonstration. En notant Ω l'ensemble des événements, l'espérance de Y est par définition $E(Y) = \sum_{\omega\in\Omega} Y(\omega)P(\omega)$, où $P(\omega)$est la probabilité de l'événement ω. On a donc

$$E(aY+bZ) = \sum_{\omega\in\Omega}\big[aY(\omega)+bZ(\omega)\big]P(\omega) = a\sum_{\omega\in\Omega}Y(\omega)P(\omega) + b\sum_{\omega\in\Omega}Z(\omega)P(\omega) = aE(Y)+bE(Z)$$

En posant $m = E(Y)$, on a $(Y-m)^2 = Y^2 - 2mY + m^2$ et en prenant l'espérance, il vient $V(Y) = E\big[(Y-m)^2\big] = E(Y^2) - 2mE(Y) + m^2 = E(Y^2) - m^2$, ce qui est l'égalité (ii). Enfin, on a $E(aY+b) = am+b$ et $(aY+b) - E(aY+b) = a(Y-m)$, d'où $V(aY+b) = E\big[(a^2(Y-m)^2\big] = a^2E\big[(Y-m)^2\big] = a^2V(Y)$. ■

2. Permutations

Nous allons étudier plus précisément les transformations bijectives d'un ensemble fini.

Définition

Une transformation bijective d'un ensemble fini E s'appelle une *permutation de E*. On note $\mathscr{S}(E)$ l'ensemble des permutations de E et si n est un entier au moins égal à 2, on note $\mathscr{S}_n$ l'ensemble des permutations de $\{1, 2, \ldots, n\}$. L'ensemble $\mathscr{S}_n$ est un groupe de transformations appelé *groupe des permutations*.

Nous avons montré page 60 que si E possède n éléments, alors le groupe $\mathscr{S}(E)$ possède $n!$ éléments.

Pour étudier les permutations d'un ensemble E à n éléments, il suffit évidemment de considérer les permutations de l'ensemble $\{1, \ldots, n\}$. Supposons donc désormais $E = \{1, \ldots, n\}$, où $n \geqslant 2$, et commençons par étudier les permutations les plus simples.

2.1 Les cycles

Soient p un entier tel que $2 \leqslant p \leqslant n$ et $a_1, a_2, \ldots, a_p$ des entiers deux à deux différents appartenant à l'ensemble $E = \{1, \ldots, n\}$. Définissons une permutation c de E en posant

$$c(a_i) = a_{i+1} \text{ si } 1 \leqslant i \leqslant p-1 \ , \quad c(a_p) = a_1 \quad \text{et} \quad c(k) = k \text{ si } k \notin \{a_1, \ldots, a_p\}.$$

Par définition, la transformation c permute circulairement les entiers $a_1, \ldots, a_p$:

$$c \ : \quad a_1 \mapsto a_2 \mapsto \cdots \mapsto a_{p-1} \mapsto a_p \mapsto a_1$$

et laisse fixe tous les autres entiers.

Définitions

La permutation c définie ci-dessus s'appelle un *p-cycle* et se note $c=(a_1\ a_2\ \cdots\ a_p)$. L'ensemble $\{a_1, \ldots, a_p\}$ est le *support* du cycle c et l'entier p est la *longueur* du cycle. Un 2-cycle s'appelle une *transposition*.

Exemples 1

- Soient a et b deux éléments différents de $\{1, 2, \ldots, n\}$. La transposition $c = (a\ b)$ échange les entiers a et b en laissant fixes tous les autres entiers. On a $c^2 = c \circ c = \mathrm{id}_E$. La transposition $(b\ a)$ échange aussi les entiers a et b en laissant fixes les autres, donc on a $(a\ b) = (b\ a)$.
- La permutation $c = (3\ 2\ 5)$ est un 3-cycle de $\mathscr{S}_6$: il est défini par $c(1)=1$, $c(2)=5$, $c(3)=2$, $c(4)=4$, $c(5)=3$ et $c(6)=6$. Remarquons que l'on a aussi $c=(2\ 5\ 3)=(5\ 3\ 2)$. Puisqu'on a $c^2(3)=c\circ c(3)=c(2)=5$, $c^2(2)=c(5)=3$, $c^2(5)=c(3)=2$, la permutation c^2 est déterminée par

$$c^2 \ : \quad 3 \mapsto 5 \mapsto 2 \mapsto 3 \qquad \text{et } c^2(k) = k \text{ si } k \in \{1, 4, 6\}\,;$$

  cela montre que c^2 est le 3-cycle $(3\ 5\ 2)$.
  On a aussi $c^3(3) = c(5) = 3$, $c^3(2) = c(3) = 2$, $c^3(5) = c(2) = 5$ et $c^3(k) = k$ si $k \in \{1, 4, 6\}$, donc $c^3 = \mathrm{id}_E$. Cette égalité s'écrit $c \circ c^2 = \mathrm{id}_E$, donc il vient $c^{-1} = c^2$.

Exemple 2. L'inverse du p-cycle $(a_1\ a_2\ \cdots\ a_p)$ est le p-cycle $(a_p\ \cdots\ a_2\ a_1)$.

Soit $c = (a_1\ a_2\ \cdots\ a_p)$ un p-cycle. Les itérés de a_1 par c sont successivement $a_2, \ldots, a_p, a_1, a_2, \ldots$ et plus précisément, nous avons $c^i(a_1) = a_{i+1}$ pour $1 \leqslant i \leqslant p-1$ et $c^p(a_1) = a_1$. Si k est un entier compris entre 1 et p, alors $a_k = c^{k-1}(a_1)$, donc

$$c^p(a_k) = c^p \circ c^{k-1}(a_1) = c^{p+k-1}(a_1) = c^{k-1} \circ c^p(a_1) = c^{k-1}(a_1) = a_k\,.$$

Ainsi la permutation c^p laisse fixe les entiers a_k, et comme elle laisse fixe les autres entiers, on en déduit que c^p est l'identité. Il s'ensuit $c^{-1} = c^{p-1} = (a_p\ \cdots\ a_2\ a_1)$.

Proposition. *Si $c = (a_1\ a_2\ \cdots\ a_p)$ est un p-cycle appartenant à $\mathscr{S}_n$, alors $c^p = \mathrm{id}_E$ et p est le plus petit entier $k > 0$ tel que $c^k = \mathrm{id}_E$.*

Composés de permutations

Notation. Si s et s' sont des permutations de E, la composée $s \circ s'$ se note simplement ss'.

Exemple 3. Dans le groupe $\mathscr{S}_6$, posons $s = (2\ 3\ 4)$ et $s' = (1\ 2)$. La composée ss' est déterminée par

$$ss' \ :\ \begin{cases} 1 \overset{s'}{\longmapsto} 2 \overset{s}{\longmapsto} 3 \\ 3 \overset{s'}{\longmapsto} 3 \overset{s}{\longmapsto} 4 \\ 4 \overset{s'}{\longmapsto} 4 \overset{s}{\longmapsto} 2 \\ 2 \overset{s'}{\longmapsto} 1 \overset{s}{\longmapsto} 1 \end{cases} \qquad \text{et} \quad ss'(k) = k \text{ si } k \notin \{1,2,3,4\}.$$

On a donc $ss' = (1\ 3\ 4\ 2)$.

Exemple 4. Soit $c = (a_1\ a_2\ \cdots\ a_p)$ un p-cycle appartenant à $\mathscr{S}_n$ et soit s une permutation de $\{1, 2, \ldots, n\}$. Montrons que la permutation $sc(s^{-1})$ est égale au p-cycle $\big(s(a_1)\ s(a_2)\ \cdots\ s(a_p)\big)$.

La permutation $c' = sc(s^{-1})$ se déduit de c par le changement de référentiel s (page 24). Elle est définie par les relations

$$c'\big(s(a_i)\big) = s\big(c(a_i)\big) = \begin{cases} s(a_{i+1}) & \text{si } 1 \leqslant i < p \\ s(a_1) & \text{si } i = p, \end{cases}$$

donc c' permute circulairement les entiers $s(a_1), s(a_2), \ldots, s(a_p)$ et laisse fixe les autres.

Remarque

Soit s une permutation de E. L'opération qui transforme une permutation f de E en sfs^{-1} est un changement de référentiel très utilisé dans les calculs. On a notamment la propriété suivante :

si f et g sont des permutations de E, alors $s(fg)s^{-1} = (sfs^{-1})(sgs^{-1})$.

En effet, on a $(sfs^{-1})(sgs^{-1}) = sf(s^{-1}s)gs^{-1} = sf\,\mathrm{id}_E\,gs^{-1} = s(fg)s^{-1}$.

En général, pour calculer le composé ss' de deux permutations, il faut tenir compte de l'ordre des facteurs : pour tout élément i, on obtient $(ss')(i) = s \circ s'(i)$ en transformant d'abord i par s', puis en appliquant s. Voici un cas important où l'ordre des facteurs n'a pas d'importance.

Proposition. *Si c_1 et c_2 sont des cycles dont les supports n'ont aucun élément en commun, alors $c_1c_2 = c_2c_1$.*

Démonstration. Posons $c_1 = (a_1\ a_2\ \cdots\ a_p)$. D'après l'exemple 4, nous avons l'égalité $c_2c_1c_2^{-1} = \big(c_2(a_1)\ c_2(a_2)\ \ldots\ c_2(a_p)\big)$. Par hypothèse, aucun des a_i n'est dans le support de c_2, donc $c_2(a_i) = a_i$. Il vient donc $c_2c_1c_2^{-1} = c_1$, d'où $c_2c_1 = c_1c_2$ en composant à droite avec c_2. ■

2.2 Décomposition en cycles

Soit s une permutation de $E = \{1, 2, \ldots, n\}$ et soit $a_0 \in E$.

Pour tout entier $i \geqslant 0$, posons $a_i = s^i(a_0)$, de sorte que les itérés de a_0 sont $a_0, a_1, a_2, \ldots$.

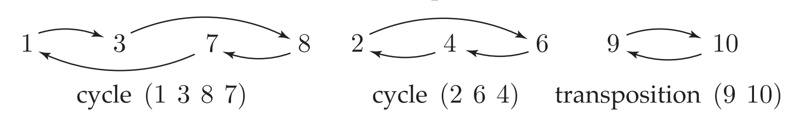
Puisque ces itérés appartiennent à l'ensemble fini $\{1, 2, \ldots, n\}$, on peut trouver deux

entiers i et j tels que $a_i = a_j$ et $i < j$. Puisque $a_i = s^i(a_0)$ et $a_j = s^j(a_0)$, on a $s^i(a_0) = s^j(a_0)$, donc $a_0 = s^{-i}\big(s^j(a_0)\big) = s^{j-i}(a_0)$ et $j - i > 0$.
On en déduit qu'il existe un plus petit entier $p > 0$ tel que $a_p = a_0$. Si $p = 1$, alors $s(a_0) = a_0$ et a_0 est un point fixe de s.
Supposons $p \geqslant 2$. Les p premiers itérés de a_0 sont $a_0, a_1, \dots, a_{p-1}$, et comme on a $a_p = a_0$, il vient $a_{p+i} = s^{p+i}(a_0) = s^i\big(s^p(a_0)\big) = s^i(a_p) = s^i(a_0) = a_i$ pour tout entier i.
La suite $a_0, a_1, a_2, \dots$ des itérés de a_0 est donc périodique de période p. La permutation s transforme les entiers $a_0, a_1, \dots, a_{p-1}$ en envoyant chacun des $p-1$ premiers sur le suivant et en envoyant a_{p-1} sur $a_p = a_0$. Les entiers $a_0, \dots, a_{p-1}$ sont donc transformés par s exactement comme par le p-cycle $(a_0\ a_1\ \cdots\ a_{p-1})$.
Remarquons que si l'on itère a_0 par s^{-1}, on obtient successivement $a_{p-1}, \dots, a_1, a_0$: l'ensemble $\mathrm{iter}(a) = \{a_0, a_1, \dots, a_{p-1}\}$ des itérés de a_0 est donc aussi l'ensemble de tous les éléments $s^k(a_0)$, où k parcourt $\mathbb{Z}$.

Exemple. Soit s la permutation de $\{1,2,3,4,5,6,7,8,9,10\}$ définie par le tableau ci-contre.
Dans les graphiques ci-dessous, les flèches permettent de visualiser les itérés d'un élément :

- les itérés de 1 sont 1, $s(1) = 3$, $s(3) = 8$ et $s(8) = 7$;
- les itérés de 2 sont : 2, $s(2) = 6$ et $s(6) = 4$;
- les éléments 9 et 10 sont échangés ;
- on a $s(5) = 5$, autrement dit 5 est un point fixe de s.

i	$s(i)$
1	3
2	6
3	8
4	2
5	5
6	4
7	1
8	7
9	10
10	9



cycle $(1\ 3\ 8\ 7)$ cycle $(2\ 6\ 4)$ transposition $(9\ 10)$

Reprenons le cas général d'une permutation s de E. Nous avons montré que l'ensemble des itérés d'un élément $a \in E$ est aussi $\mathrm{iter}(a) = \{s^k(a) \mid k \in \mathbb{Z}\}$.
Si $b \in E$, on a l'équivalence $b = s^k(a) \Longleftrightarrow a = s^{-k}(b)$: b est donc un itéré de a si et seulement si a est un itéré de b. Si b est un itéré de a, alors tout itéré de b est un itéré de a et réciproquement, donc les éléments a et b ont même ensemble d'itérés.
On en déduit que si des ensembles $\mathrm{iter}(a)$ et $\mathrm{iter}(a')$ ont en commun un élément x, alors ces ensembles sont égaux. En effet, x étant un itéré de a, on a $\mathrm{iter}(x) = \mathrm{iter}(a)$, et de même on a $\mathrm{iter}(x) = \mathrm{iter}(a')$; il s'ensuit $\mathrm{iter}(a) = \mathrm{iter}(a')$.
Les différents ensembles $\mathrm{iter}(a), \mathrm{iter}(b), \dots$ sont donc des parties de E deux à deux disjointes et leur réunion est E.

Rappelons ce que nous avons montré avant l'exemple : si un élément a possède q itérés, où $q \geqslant 2$, alors pour tout $u \in \mathrm{iter}(a)$, on a $\mathrm{iter}(a) = \{u, s(u), \dots, s^{q-1}(u)\}$ et tous les éléments de $\mathrm{iter}(a)$ sont transformés par s selon le q-cycle $c = \big(u\ s(u)\ \cdots\ s^{q-1}(u)\big)$.

Théorème. *Soit s une permutation de E différente de l'identité.*

- *Il existe des cycles $c_1, \dots, c_m$, dont les supports sont des parties deux à deux disjointes et tels que $s = c_1 c_2 \cdots c_m$.*

► *Le plus petit entier $r > 0$ tel que $s^r = \mathrm{id}_E$ est le ppcm des longueurs des cycles $c_1, \ldots, c_m$.*

Comme les supports des cycles c_i sont des parties deux à deux disjointes, on peut composer ces cycles dans l'ordre qu'on veut (proposition page 72).
Il s'ensuit que l'on a $s^k = c_1^k c_2^k \cdots c_m^k$ pour tout entier k.

Démonstration. Les éléments de E se répartissent en les parties $C_1 = \mathrm{iter}(a_1), C_2 = \mathrm{iter}(a_2), \ldots, C_m = \mathrm{iter}(a_m)$ deux à deux disjointes. Quand on itère s, tous les éléments de C_k décrivent un même cycle c_k de support C_k. Soit $x \in C_k$. On a donc $c_k(x) = s(x)$ et $c_k(x) \in C_k$. Pour tout $j \neq k$, ni x, ni $c_k(x)$ n'appartiennent à C_j, donc $c_j(x) = x$ et $c_j c_k(x) = c_k(x)$. Il vient donc $c_1 c_2 \cdots c_m(x) = c_k(x) = s(x)$. Cette égalité étant vraie pour tout $x \in C_k$ et pour tout k, elle est vraie quel que soit $x \in E$.
Si x est un élément de C_k, il revient au même d'itérer x par s ou par c_k, donc $s^r(x) = c_k^r(x)$ pour tout entier r. Pour que l'on ait $s^r(x) = x$ quel que soit $x \in C_k$, il faut et il suffit que r soit multiple de la longueur de c_k. Pour que l'on ait $s^r(x) = x$ quel que soit $x \in E$, une condition nécessaire et suffisante est donc que r soit multiple du ppcm des longueurs des cycles c_k. ■

Exemple. La permutation s de l'exemple précédent se décompose en trois cycles : le 4-cycle $(1\ 3\ 8\ 7)$, le 3-cycle $(2\ 6\ 4)$ et le 2-cycle $(9\ 10)$. On a $s = (1\ 3\ 8\ 7)(2\ 6\ 4)(9\ 10)$. Le plus petit entier r tel que $s^r = \mathrm{id}$ est $r = \mathrm{ppcm}(4, 3, 2) = 12$.

Algorithme de décomposition en cycles

Soit s une permutation de $\{1, 2, \ldots, n\}$. On suppose s différent de l'identité. L'algorithme suivant produit la liste L des cycles composant s. Un cycle $(a_0\ a_1\ \cdots\ a_p)$ est représenté par la liste $[a_0, \ldots, a_p]$. On commence par chercher le cycle décrit par les itérés de 1 en marquant chacun des entiers obtenus ; puis on cherche les itérés du plus petit entier non marqué et l'on continue ainsi jusqu'à ce que tous les entiers soient marqués. On calcule chaque fois le nombre ℓ d'itérés, car si l'on trouve un point fixe de s, on le supprime puisqu'il ne produit pas de cycle.

initialisations : $(L, C \leftarrow$ listes vides$)$ $(m[i] = 0$ pour tout i tel que $1 \leqslant i \leqslant n)$
les entiers marqués seront les entiers i tels que $m(i) = 1$.

a) $u \leftarrow \min\{i \mid 1 \leqslant i \leqslant n$ et $m[i] = 0\}$;

b) $C \leftarrow C, u$ (*on ajoute u à la liste C*) ; $m[u] \leftarrow 1$; $\ell \leftarrow 1$; $v \leftarrow s(u)$;

c) tant que $v \neq u$:

   $C \leftarrow C, v$ (*on ajoute v à la liste C*) ; $m[v] \leftarrow 1$; $\ell \leftarrow \ell + 1$; $v \leftarrow s(v)$;

d) si $\ell \geqslant 2$, $L \leftarrow L, C$ (*on ajoute la liste C à la liste L*) ;

e) aller en a).

À la fin de l'algorithme, on obtient une liste L de listes $C_1, C_2, \ldots$ La permutation s est la composée des cycles représentés par les listes C_i.

Groupe engendré par une permutation

Soit s une permutation de l'ensemble $E=\{1,2,\ldots,n\}$, différente de l'identité. Rappelons que l'ensemble $T=\{s^k\mid k\in\mathbb{Z}\}$ est un groupe de transformations de E (page 28). D'après le théorème, il y a un plus petit entier $r>0$ tel que $s^r=\mathrm{id}_E$. Il s'ensuit que les transformations $\mathrm{id}_E=s^0, s, \ldots, s^{r-1}$ sont deux à deux différentes et que $s^{r+j}=s^j$ pour tout entier j. On a donc simplement $T=\{\mathrm{id}_E, s, \ldots, s^{r-1}\}$ et le groupe T possède r éléments.
Soient s^i et s^j des éléments de T, donc $s^is^j=s^{i+j}$. En appelant k le reste de la division de $i+j$ par r, il vient $i+j=rq+k$ et $s^is^j=s^{rq+k}=(s^r)^qs^k=(\mathrm{id}_E)^qs^k=s^k$, où k est compris entre 0 et r. Cela permet de calculer dans le groupe T. On détermine r au moyen de la décomposition en cycles et du théorème précédent.

Exemples

- Pour la permutation s de l'exemple page 73, on a $T=\{s^i \mid 0\leqslant i\leqslant 11\}$ (avec la convention $s^0=\mathrm{id}_E$).
- Supposons que s est un p-cycle $(a_1\ a_2\ \ldots a_p)$. Alors on a $T=\{\mathrm{id}_E, s, \cdots, s^{p-1}\}$. Dans le groupe T engendré par s, les règles de calcul sont les mêmes que dans le groupe engendré par une rotation d'angle $2\pi/p$ (exemple 3 page 28).

2.3 Puissance d'un cycle

Les permutations les plus utilisées en pratique sont les cycles. En particulier, on est souvent amené à étudier les puissances c^k d'un cycle c.
Rappelons que si p et q sont des entiers positifs, on a la relation

$$\mathrm{ppcm}(p,q)\times\mathrm{pgcd}(p,q)=pq$$

Proposition. *Soient c un p-cycle et k un entier strictement positif et non multiple de p. Posons $d=\mathrm{pgcd}(p,k)$. La permutation c^k est composée de d cycles, tous de même longueur p/d. Pour que c^k soit un cycle, il faut et il suffit que p et k soient premiers entre eux.*

Démonstration. Posons $c=(a_0\ a_1\ \cdots\ a_p)$ et $s=c^k$. Soit $x\in\{a_0,\ldots,a_p\}$. Le cycle des itérés de x par s a pour longueur le plus petit entier $m>0$ tel que $s^m(x)=x$, c'est-à-dire $c^{km}(x)=x$. Puisque c est un p-cycle, l'égalité $c^i(x)=x$ se produit si et seulement si i est multiple de p. On en déduit

$$\begin{aligned} s^m(x)=x \iff c^{km}(x)=x &\iff km \text{ est multiple de } p\\ &\iff km \text{ est multiple de } p \text{ et de } k\\ &\iff km \text{ est multiple de } \mathrm{ppcm}(p,k)=\frac{pk}{d}\\ &\iff m \text{ est multiple de } \frac{p}{d}.\end{aligned}$$

Ainsi, lorsqu'on décompose c^k en cycles, tous les cycles obtenus ont même longueur p/d ; il y a donc d cycles. Pour que c^k soit un cycle, il faut et il suffit que $d=1$, c'est-à-dire que p et k soient premiers entre eux. ■

Exemples

► Soit le 12-cycle $c = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12)$. On a $c^{12} = \mathrm{id}$, donc $c^{11} = c^{-1}$,

$$\begin{aligned}
&c^2 = (1\ 3\ 5\ 7\ 9\ 11)(2\ 4\ 6\ 8\ 10\ 12) \quad \text{et} \quad c^{10} = (c^2)^{-1}\\
&c^3 = (1\ 4\ 7\ 10)(2\ 5\ 8\ 11)(3\ 6\ 9\ 12) \quad \text{et} \quad c^9 = (c^3)^{-1}\\
&c^4 = (1\ 5\ 9)(2\ 6\ 10)(3\ 7\ 11)(4\ 8\ 12) \quad \text{et} \quad c^8 = (c^4)^{-1}\\
&c^5 = (1\ 6\ 11\ 4\ 9\ 2\ 7\ 12\ 5\ 10\ 3\ 8) \qquad \text{et} \quad c^7 = (c^5)^{-1}\\
&c^6 = (1\ 7)(2\ 8)(3\ 9)(4\ 10)(5\ 11)(6\ 12)\,.
\end{aligned}$$

Les entiers positifs compris entre 1 et 12 et premiers à 12 sont 1, 5, 7 et 11 : si k un entier compris entre 1 et 11, la permutation c^k n'est un cycle que si $k \in \{1, 5, 7, 11\}$. Si k est l'un des entiers 2, 3, 4, 6, 8, 9 ou 10, la permutation c^k se décompose en deux, trois, quatre ou six cycles de longueur 6, 4, 3 ou 2.

► Soit p un nombre premier, c'est-à-dire que p est un entier au moins égal à 2 dont les seuls diviseurs positifs sont 1 et p. Si c est un p-cycle, alors pour tout entier k non multiple de p, c^k est un p-cycle. En effet, 1 est le seul diviseur commun à p et k.

2.4 Le rôle des transpositions

Rappelons qu'une transposition est par définition un 2-cycle $(a\ b)$.

a) *Si des éléments sont écrits dans un ordre, on peut en échanger deux quelconques par une succession d'échanges entre éléments consécutifs.*

Par exemple, pour échanger 1 et 4 dans la succession $1, 2, 3, 4$, nous pouvons effectuer les échanges suivants :

échange	1	2	3	4
(1 2)	**2**	**1**	3	4
(1 3)	2	**3**	**1**	4
(1 4)	2	3	**4**	**1**
(4 3)	2	**4**	**3**	1
(4 2)	**4**	**2**	3	1

On a l'égalité $(1\ 4) = (4\ 2)(4\ 3)(1\ 4)(1\ 3)(1\ 2)$ et dans cette composée, chaque transposition échange deux éléments qui étaient consécutifs.

b) *Tout cycle est un composé de transpositions.*

On a en effet l'égalité $(a_1\ a_2\ \cdots\ a_p) = (a_1\ a_2)(a_2\ a_3)\cdots(a_{p-1}\ a_p)$.

Nous avons montré que toute permutation est une composée de cycles (l'identité étant égale à la composée $\tau\tau$, où τ est une transposition quelconque). On en déduit :

Proposition. *Toute permutation est une composée de transpositions.*

Puisqu'une transposition peut toujours se réaliser par échanges d'éléments consécutifs, il en va de même pour une permutation quelconque. Voici une méthode de tri fondée sur cette propriété.

Le tri à bulles

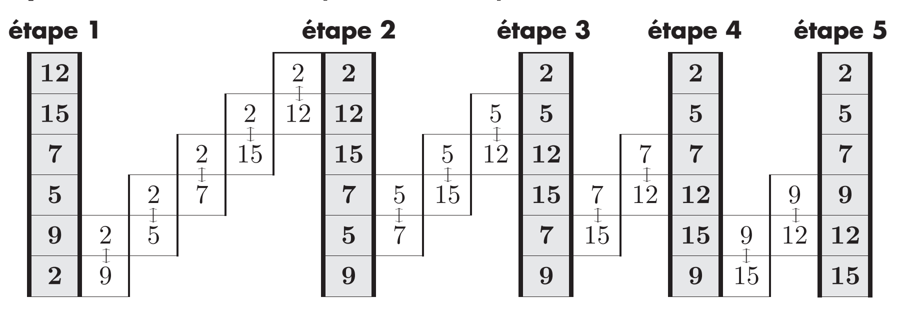
On dispose de données $a_1, a_2, \ldots, a_n$ deux à deux comparables que l'on veut classer par exemple par ordre croissant ; ces données peuvent être des nombres, ou encore des chaînes de caractères qu'on classera selon leur longueur.
Notons $a \prec b$ la propriété « a est plus petit ou égal à b ». Cette relation doit être une relation d'ordre, c'est-à-dire que l'on a :

$$a \prec a \ , \quad (a \prec b \text{ et } b \prec c) \Longrightarrow a \prec c \ , \quad (a \prec b \text{ et } b \prec a) \Longrightarrow a = b.$$

Initialement, les données sont présentées en désordre dans une liste ou un tableau.

Principe du tri à bulles. Il consiste à parcourir la liste en commençant par la fin et en effectuant un échange à chaque fois que l'on trouve deux éléments successifs qui ne sont pas dans le bon ordre.

Exemple. La liste à trier est $[12, 15, 7, 5, 9, 2]$.



étape 1 : le nombre 2 étant le plus petit, il est successivement échangé avec tous les éléments de la liste.
étape 2 : le nombre 5 est échangé successivement avec 7, 15 et 12 qui sont supérieurs.
étape 3 : on échange 7 avec 15, puis avec 12.
étape 4 : maintenant, 15 précède 9, donc on les échange, puis on échange 9 et l'élément 12 qui le précède.
étape 5 : il n'y a plus d'échange possible, donc la liste est triée selon l'ordre croissant.

Cet algorithme doit son nom au fait que les éléments « les plus légers remontent » vers le haut de la liste, comme des bulles de gaz dans un liquide. S'il y a n données à trier, le nombre d'échanges à effectuer est au maximum $n-1$ à la première étape, $n-2$ à la deuxième, etc ; le nombre maximum d'échanges est donc $1+2+\cdots+(n-1)=n(n-1)/2$. Puisqu'un échange $a \leftrightarrow b$ se réalise par une succession $(c \leftarrow a), (a \leftarrow b), (b \leftarrow c)$ de trois affectations, il faut au plus $3n(n-1)/2$ opérations pour trier n données.

Algorithme du tri à bulles

initialisations : tableau de données $D[i]$ $1 \leqslant i \leqslant n$ à trier selon l'ordre croissant ; $i \leftarrow 1$;

tant que $i < n$, faire :

i) pour j de n à $i+1$: si $D[j] < D[j-1]$, échanger $D[j]$ et $D[j-1]$;

ii) $i \leftarrow i+1$.

3. Graphes

Définitions

Un *graphe* est la donnée d'un ensemble S, fini et non vide, de *sommets* et d'un ensemble A d'*arcs* : un arc est une paire $\{a,b\}$ de sommets. Si $\{a,b\}$ est un arc, on dit que les sommets a et b sont *adjacents* ou *joints par un arc* et l'on note $\overset{\frown}{a,b}$ ou $\overset{\frown}{b,a}$ l'arc joignant ces sommets.

On peut visualiser un graphe par un dessin : les sommets sont représentés par des points et si deux sommets sont adjacents, on les relie par une ligne.
Ci-contre le graphe de sommets $S = \{a,b,c,d,e\}$ ayant pour arcs $\overset{\frown}{a,b}$, $\overset{\frown}{a,c}$, $\overset{\frown}{b,c}$, $\overset{\frown}{c,d}$, $\overset{\frown}{d,e}$, $\overset{\frown}{b,d}$. Les sommets b et c sont adjacents, de même que c et d, mais les sommets a et d ne le sont pas, c et e non plus.

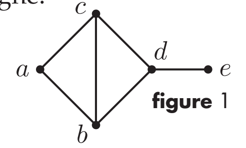


figure 1

Exemple. Les liaisons aériennes assurées par une compagnie définissent un graphe : les sommets représentent les villes desservies et deux sommets sont adjacents s'il y a une liaison entre les villes correspondantes.

Le *degré* d'un sommet est le nombre d'arcs passant par ce sommet. Puisque chaque arc passe par exactement deux sommets, la somme des degrés vaut deux fois le nombre d'arcs.

Définitions

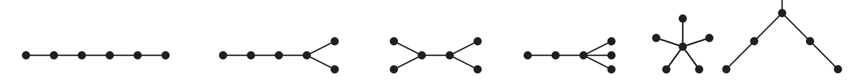
Soit G un graphe.

- Un *chemin* est une suite $s_1, s_2, \ldots, s_n$ de sommets adjacents deux à deux différents ; le premier sommet est l'origine du chemin et le dernier est l'extrémité.
- Le graphe G est *connexe* si deux sommets différents peuvent toujours être reliés par un chemin.
- Soit $s_1, s_2, \ldots, s_n$ un chemin. Si $n \geqslant 3$ et si s_n est adjacent à s_1, on dit que $(s_1, s_2, \ldots, s_n, s_1)$ est un *cycle*.

Exemple. Dans le graphe de la figure 1, (a,b,c,d,e) et (a,b,d,c) sont des chemins, (a,c,d,b,a) est un cycle ; le graphe est connexe.

Définition

Un *arbre* est un graphe connexe qui n'a pas de cycle.



les arbres à six sommets

Caractérisation d'un arbre. *Soit G un graphe ayant n sommets. Les propriétés suivantes sont équivalentes :*

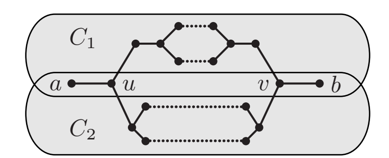
i) *G est un arbre ;*

ii) *pour tous sommets a,b différents, il existe un unique chemin d'origine a et d'extrémité b ;*

iii) *G est connexe et possède $n-1$ arcs.*

Démonstration.

Il suffit de raisonner dans le cas $n \geqslant 3$. Supposons que G est un arbre. Alors G est connexe donc entre deux sommets a et b différents, il existe au moins un chemin d'origine a et d'extrémité b ; supposons qu'il existe deux tels chemins C_1 et C_2 ; considérons le premier sommet u de C_1 qui est aussi sur C_2 mais dont le successeur sur C_1 n'est pas sur C_2 (u existe car $C_1 \neq C_2$). Soit v le sommet suivant sur C_1 qui soit aussi sur C_2. Les parties de C_1 et C_2 entre u et v forment alors un cycle, contrairement à l'hypothèse que G est un arbre.



Cela montre que (i) implique (ii). Réciproquement, si G possède la propriété (ii), il est évidemment connexe et sans cycle. On a donc équivalence entre (i) et (ii).

Pour démontrer que (i) implique (iii), raisonnons par récurrence sur le nombre de sommets. On peut trouver dans G un chemin $s_1, s_2, \ldots, s_n$ de longueur maximum. Supposons que a est un sommet adjacent à s_n ; si a était différent de tous les s_i, on pourrait prolonger le chemin en ajoutant a ; si a était l'un des sommets $s_1, \ldots, s_{n-2}$, le graphe contiendrait un cycle ; le seul sommet adjacent à s_n est donc s_{n-1}. En supprimant le sommet s_n et l'arc $\widehat{s_{n-1}, s_n}$, on obtient un graphe G' ayant $n-1$ sommets et $p-1$ arcs, où p est le nombre d'arcs de G. Le graphe G' n'a pas de cycle et est encore connexe, donc G' est un arbre. Par hypothèse de récurrence, G' possède $n-2$ arcs, donc on a l'égalité $p-1 = n-2$, ou encore $p = n-1$.

Il reste à montrer que (iii) implique (i). On raisonne à nouveau par récurrence sur le nombre de sommets. Supposons que G est connexe et possède $n-1$ arcs. Puisque G est connexe, les sommets ont des degrés $d_1, \ldots, d_n$ strictement positifs. L'égalité $d_1 + d_2 + \cdots + d_n = 2(n-1)$ montre que les degrés ne peuvent pas être tous strictement supérieurs à 1, donc il existe au moins un sommet a de degré 1. En supprimant de G le sommet a et l'unique arc qui y passe, on obtient un graphe G' ayant $n-1$ sommets et $n-2$ arcs. Ce graphe G' est encore connexe, il a un arc de moins que le nombre de sommets, donc c'est un arbre, par hypothèse de récurrence. Un éventuel cycle de G doit passer par a, ce qui n'est pas possible car le degré de a est égal à 1. Le graphe G n'a donc pas de cycle. ■

On a souvent besoin de pondérer les arcs : par exemple, si les arcs d'un graphe représentent des liaisons aériennes, on pourra attribuer à chaque arc le temps ou le coût de transport correspondant.

Définition
Un graphe est *pondéré* si l'on a associé à chaque arc $\widehat{a,b}$ un nombre réel $w(a,b)$ appelé le *poids* de l'arc. Le poids d'un chemin, ou d'un graphe, est la somme des poids des arcs qui le composent.

Nous allons présenter trois problèmes classiques et leur algorithme de résolution.

3.1 Arbre de recouvrement de poids minimal

Exemple. À partir d'une localité p_1 déjà raccordée au réseau du gaz, on veut alimenter les localités $p_2, p_3, \ldots, p_6$. Le plan de distribution doit minimiser la longueur de conduite à poser. Ci-contre le tableau des distances entre localités.
Formons le graphe G de sommets $1, \ldots, 6$ où deux sommets quelconques sont toujours joints par un arc. L'arc joignant i et j est pondéré par la distance $d_{i,j}$ des localités p_i et p_j.
Un plan de distribution sera représenté par un sous-graphe T ayant les propriétés suivantes :

	p_1	p_2	p_3	p_4	p_5
p_2	5				
p_3	3	4			
p_4	7	1	8		
p_5	10	6	2	7	
p_6	2	5	6	4	3

a) T doit contenir tous les sommets $1, 2, \ldots, 6$;
b) T n'a pas de cycle et entre deux sommets, il doit exister un chemin : autrement dit, T est un arbre ;
c) parmi les arbres ayant les propriétés précédentes, T doit être de poids minimal.

Un sous-graphe T ayant les propriétés (a) et (b) s'appelle un *arbre de recouvrement* de G. On va construire un arbre de recouvrement de poids minimal en sélectionnant au fur et à mesure ses sommets et ses arcs.

I) Choisissons dans G un arc de poids minimal : l'arc $\widehat{2,4}$, de poids 1, convient. Les sommets 2 et 4, avec l'arc qui les joint, forment un arbre T_1.
II) Considérons tous les arcs joignant l'un des sommets 2 ou 4 à un sommet qui n'est pas dans T_1, c'est-à-dire les arcs $\widehat{1,2}$, $\widehat{1,4}$, $\widehat{3,2}$, $\widehat{3,4}$, $\widehat{5,2}$, $\widehat{5,4}$, $\widehat{6,2}$ et $\widehat{6,4}$. Les poids sont

$d_{1,2}$	$d_{1,4}$	$d_{3,2}$	$d_{3,4}$	$d_{5,2}$	$d_{5,4}$	$d_{6,2}$	$d_{6,4}$
5	7	4	8	6	7	5	4

Parmi ces arcs, l'arc $\widehat{3,2}$ est de poids minimal $d_{3,2} = 4$. On l'ajoute à T_1 pour former l'arbre T_2 de sommets $\{2,4,3\}$ ayant pour arcs $\widehat{4,2}$ et $\widehat{2,3}$.
III) Recommençons comme à l'étape précédente : voici les poids des arcs joignant l'un des sommets de T_2 à un sommet qui n'est pas dans T_2, c'est-à-dire à 1, 5 ou 6 :

$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{5,2}$	$d_{5,3}$	$d_{5,4}$	$d_{6,2}$	$d_{6,3}$	$d_{6,4}$
5	3	7	6	2	7	5	6	4

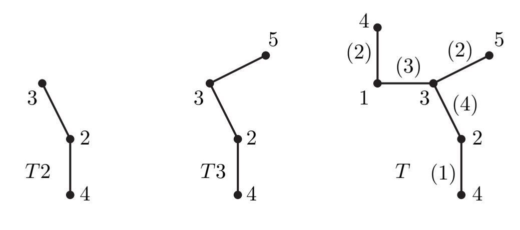
Le poids minimal est obtenu pour l'arc $\widehat{5,3}$ de poids 2 : en ajoutant cet arc à T_2, on obtient l'arbre T_3 qui a pour sommets $4, 2, 3, 5$.

IV) Continuons de même : voici les poids des arcs joignant un sommet de T_3 à l'un des sommets 1 ou 6 :

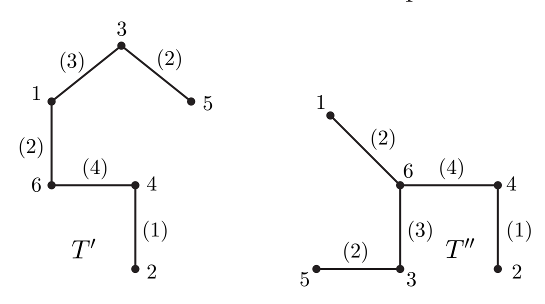
$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{1,5}$	$d_{6,2}$	$d_{6,3}$	$d_{6,4}$	$d_{6,5}$
5	3	7	10	5	6	4	3

Comme arc de poids minimal, choisissons l'arc $\widehat{1,3}$ de poids 3. En ajoutant cet arc à T_3, on obtient l'arbre T_4 de sommets $4, 2, 3, 5, 1$.

V) Il reste à sélectionner un arc de poids minimal parmi ceux qui joignent le sommet 6 à un autre. C'est l'arc $\widehat{6,1}$, de poids 2, qui convient : en l'ajoutant à T_4, on obtient un arbre T solution : il contient tous les sommets et est de poids minimal 12.



Dans cet exemple, nous aurions pu, en II, choisir l'arc $\widehat{4,6}$; cela aurait conduit à l'un des arbres T' ou T'' différents de T mais de même poids total 12.



Construction d'un arbre de recouvrement de poids minimal

Soit G un graphe pondéré connexe. En numérotant les sommets de G, on peut supposer que l'ensemble des sommets est $P = \{1, 2, \ldots, n\}$. Si i et j sont des sommets adjacents, notons $d_{i,j}$ le poids de l'arc $\widehat{i,j}$. Les sommets de l'arbre en construction seront placés dans un ensemble S et les arcs dans un ensemble A. Le symbole ∞ désigne un nombre strictement supérieur à tous les poids des arcs de G.

Étape 1. Sélectionner dans G un arc de poids minimal ; si cet arc joint les sommets a et b, on pose $S = \{a, b\}$ et $A = \left\{\widehat{a,b}\right\}$.

Étape 2. Pour tout $i \in P \setminus S$,

- s'il existe un arc joignant i à un sommet de S, trouver un sommet $k_i \in S$ tel que $d_{i,k_i} = \min\limits_{j \in S}(d_{i,j})$. Poser $\alpha_i = d_{i,k_i}$.
- s'il n'y a pas d'arc entre i et un sommet de S, poser $\alpha_i = \infty$.

Étape 3. Trouver i^* tel que $\alpha_{i^*} = \min_{j \in P \setminus S} (\alpha_j)$. Poser

$$S \leftarrow S \cup \{i^*\} \quad \text{et} \quad A \leftarrow A \cup \left\{ \widehat{i^*, k_{i^*}} \right\}.$$

Si l'ensemble S possède n éléments, alors l'arbre ayant pour arcs les éléments de A est un arbre de recouvrement de poids minimal et l'algorithme est terminé.

Étape 4. Pour tout sommet $i \in P \setminus S$ adjacent dans G à i^* et tel que $d_{i,i^*} < \alpha_i$, poser

$$\alpha_i \leftarrow d_{i,i^*} \quad \text{et} \quad k_i \leftarrow i^* ,$$

puis retourner à l'étape 3.

Remarque

À l'étape 2, on a introduit le nombre α_i dont la valeur est le plus petit poids d'un arc reliant le sommet $i \in P \setminus S$ à l'un des sommets de S. Cela permet de diminuer le nombre de comparaisons ultérieures entre poids. En effet, dans l'exemple précédent, la comparaison entre les poids $d_{1,2}$ et $d_{1,4}$ se fait en (II) : à l'issue de cette opération, on sait que le plus petit des deux poids est $d_{1,2}$; en posant $\alpha_1 = d_{1,2} = 5$, il suffira, en (III), de comparer les nombres $d_{1,3}$ et α_1 ; la valeur $k_1 = 2$ indique que c'est l'arc $\widehat{1, 2}$ qui a le plus petit des deux poids. On a de même $k_3 = 2$, $\alpha_3 = d_{3,2} = 4$, $k_5 = 2$, $\alpha_5 = d_{5,2} = 2$, $k_6 = 4$ et $\alpha_6 = d_{6,4} = 4$. Quand on ajoute à S un nouveau sommet s, on peut être amené ensuite à considérer un arc $\widehat{i, s}$ de poids inférieur à la valeur α_i : c'est pourquoi, dans l'étape 4, on donne dans ce cas à α_i la valeur $d_{i,s}$ avant de retourner à l'étape 3.

Exemples d'application. La recherche d'un arbre de recouvrement de poids minimal se rencontre dans de nombreux problèmes. Nous avons déjà mentionné l'organisation d'un service de distribution (courrier, marchandises, information) ; voici deux autres exemples.

- En Biologie, on cherche à construire des arbres phylogéniques : ces arbres décrivent les relations de parenté entre les organismes d'un groupe sélectionné et permettent de faire des hypothèses en théorie de l'évolution.
  On choisit un ensemble de caractères morphologiques ou génétiques et l'on forme un graphe pondéré : les sommets représentent les organismes considérés et le poids d'un arc entre deux organismes reflète l'écart entre leurs caractères, par exemple au moyen d'un indice fondé sur la fréquence et la stabilité de certaines mutations génétiques. Si l'on suppose que l'évolution des organismes se fait le plus probablement au moindre coût génétique, alors un arbre de recouvrement de poids minimal renseigne sur l'ordre dans lequel les mutations ont pu se produire.
- La Robotique conçoit des automates électro-mécaniques programmables pour des tâches spécifiques. Leur fonctionnement se modélise aisément par un graphe : un sommet représente un état de l'automate et un arc est une transition directe entre deux états. Pour pondérer les arcs, on peut utiliser par exemple la durée de la transition ou l'énergie consommée.

Une bonne programmation de l'automate doit lui permettre de passer au moindre coût d'un état quelconque à un autre. Chaque arbre de recouvrement de poids minimal apporte donc une réponse raisonnable à cette demande d'optimisation.

3.2 Chemin de poids minimum d'un sommet à un autre

Dans un graphe où la pondération des arcs représente par exemple des distances ou des durées d'exécution de tâches, il est naturel de chercher les chemins les plus courts possibles entre sommets.

Par exemple, en Linguistique, on définit des notions de proximité lexicale entre langues d'un groupe donné. Si l'on forme un graphe dont les sommets représentent les langues considérées et dont les arcs sont pondérés par la proximité, un chemin de poids minimal entre deux sommets décrit les étapes d'une influence linguistique possible.

Exemple. Voici une table des temps de transport routier entre cinq localités A, B, C, D, E (l'unité est le quart d'heure) ; la durée d'un trajet n'est pas forcément le même dans un sens ou dans l'autre (à cause de la topographie, par exemple) ; le signe ∞ indique qu'il n'y a pas de liaison directe entre les localités.

	A	B	C	D	E
A	0	4	1	∞	6
B	5	0	3	6	2
C	1	1	0	4	4
D	∞	2	10	0	9
E	1	1	3	7	0

Représentons ces données par un graphe à cinq sommets correspondant aux localités A, B, C, D, E ; pour simplifier, nous numérotons dans cet ordre les sommets de 1 à 5. Pour chaque paire de sommets i, j, il y a un arc muni de deux poids : le poids $d(i, j)$ est le temps de transport de i vers j et le poids $d(j, i)$ est le temps de transport de j vers i ; on donne au signe ∞ une très grande valeur, par exemple 1000.
Formons le tableau

$$D^0 = \begin{bmatrix} 0 & 4 & 1 & 1000 & 6 \\ 5 & 0 & 3 & 6 & 2 \\ 1 & 1 & 0 & 4 & 4 \\ 1000 & 2 & 10 & 0 & 9 \\ 1 & 1 & 3 & 7 & 0 \end{bmatrix}$$

Si i et j sont des entiers entre 1 et 5, on note $D^0_{i,j}$ le nombre situé à l'intersection de la i-ième ligne et de la j-ième colonne. Par exemple, $D^0_{2,4} = 6$.
On va aussi utiliser des tableaux P à cinq lignes et cinq colonnes : le nombre $P_{i,j}$, situé à l'intersection de la i-ième ligne et de la j-ième colonne de P, sera le numéro du sommet qui suit le sommet i dans un plus court chemin de i vers j déjà découvert.

Initialement, on choisit simplement le chemin (i, j) comme plus court chemin ; en notant P^0 le premier tableau, on a donc $P^0_{i,j} = j$ et

$$P^0 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

Mais pour aller par exemple de 5 à 3, il vaut mieux passer par le sommet 1, car on a $D^0_{5,1} + D^0_{1,3} = 1 + 1 < 3 = D^0_{5,3}$. On va systématiquement comparer la durée d'un trajet direct à celle d'un trajet passant par le sommet 1.

I) Formons un tableau D^1 contenant les durées des plus courts chemins passant éventuellement par le sommet 1. On pose donc

$$D^1_{i,j} = \begin{cases} D^0_{i,1} + D^0_{1,j} & \text{si } D^0_{i,1} + D^0_{1,j} < D^0_{i,j} \\ D^0_{i,j} & \text{sinon} \end{cases}$$

Si l'on a trouvé que le chemin $(i, 1, j)$ est plus rapide que (i, j), il faut modifier le tableau P en posant $P_{i,j} = 1$. On définit donc un tableau P^1 tel que

$$P^1_{i,j} = \begin{cases} 1 & \text{si } D^0_{i,1} + D^0_{1,j} < D^0_{i,j} \\ j & \text{sinon} \end{cases}$$

Nous avons remarqué que le chemin $(5, 1, 3)$ a un poids 2 inférieur au poids de $(5, 3)$, donc on pose $D^1_{5,3} = 2$ et $P^1_{5,3} = 1$. On vérifie qu'aucun autre chemin ne peut être raccourci en passant par le sommet 1. Il vient donc

$$D^1 = \begin{bmatrix} 0 & 4 & 1 & 1000 & 6 \\ 5 & 0 & 3 & 6 & 2 \\ 1 & 1 & 0 & 4 & 4 \\ 1000 & 2 & 10 & 0 & 9 \\ 1 & 1 & 2 & 7 & 0 \end{bmatrix} \qquad P^1 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 1 & 4 & 5 \end{bmatrix}$$

II) On cherche maintenant le plus court chemin de i vers j passant éventuellement par les sommets 1 et 2. Pour cela, on recommence les opérations précédentes en comparant $D^1_{i,j}$ et $D^1_{i,2} + D^1_{2,j}$ et en retenant le chemin le plus court. Voici les améliorations possibles :

$$\begin{aligned} D^1_{1,2} + D^1_{2,4} &= 4 + 6 = 10 < 1000 = D^1_{1,4} \\ D^1_{4,2} + D^1_{2,1} &= 2 + 5 = \ 7 \ < 1000 = D^1_{4,1} \\ D^1_{3,2} + D^1_{2,5} &= 1 + 2 = \ 3 \ < \ \ 4 \ \ = D^1_{3,5} \\ D^1_{4,2} + D^1_{2,1} &= 2 + 5 = \ 7 \ < 1000 = D^1_{4,1} \\ D^1_{4,2} + D^1_{2,3} &= 2 + 3 = \ 5 \ < \ \ 10 \ \ = D^1_{4,3} \\ D^1_{4,2} + D^1_{2,5} &= 2 + 2 = \ 4 \ < \ \ 9 \ \ = D^1_{4,5} \end{aligned}$$

On pose donc

$$D^2 = \begin{bmatrix} 0 & 4 & 1 & 10 & 6 \\ 5 & 0 & 3 & 6 & 2 \\ 1 & 1 & 0 & 4 & 3 \\ 7 & 2 & 5 & 0 & 4 \\ 1 & 1 & 2 & 7 & 0 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 1 & 2 & 3 & 2 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 2 \\ 2 & 2 & 2 & 4 & 2 \\ 1 & 2 & 1 & 4 & 5 \end{bmatrix}.$$

III) On compare maintenant $D^2_{i,j}$ et $D^2_{i,3} + D^2_{3,j}$. On a

$$\begin{aligned} D^2_{1,3} + D^2_{3,2} &= 1 + 1 = 2 < \ 4 \ = D^2_{1,2} \\ D^2_{1,3} + D^2_{3,4} &= 1 + 4 = 5 < 10 = D^2_{1,4} \\ D^2_{2,3} + D^2_{3,1} &= 3 + 1 = 4 < \ 5 \ = D^2_{2,1} \\ D^2_{4,3} + D^2_{3,1} &= 5 + 1 = 6 < \ 7 \ = D^2_{4,1} \\ D^2_{5,3} + D^2_{3,4} &= 2 + 4 = 6 < \ 7 \ = D^2_{5,4} \end{aligned}$$

et $D^2_{i,3} + D^2_{3,j} \geqslant D^2_{i,j}$ dans les autres cas, donc

$$D^3 = \begin{bmatrix} 0 & 2 & 1 & 5 & 4 \\ 4 & 0 & 3 & 6 & 2 \\ 1 & 1 & 0 & 4 & 3 \\ 6 & 2 & 5 & 0 & 4 \\ 1 & 1 & 2 & 6 & 0 \end{bmatrix} \qquad P^3 = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 \\ 3 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 2 \\ 2 & 2 & 2 & 4 & 2 \\ 1 & 2 & 1 & 1 & 5 \end{bmatrix}.$$

IV) Regardons si on peut raccourcir les chemins en les faisant passer par le sommet 4. On a $D^3_{i,4} + D^3_{4,j} \geqslant D^3_{i,j}$ pour tous i, j, donc

$$D^4 = D^3 \qquad \text{et} \qquad P^4 = P^3 .$$

V) Il reste à considérer le passage éventuel par le sommet 5. On a

$$\begin{aligned} D^4_{2,5} + D^4_{5,1} &= 2 + 1 = 3 < 4 = D^4_{2,1} \\ D^4_{4,5} + D^4_{5,1} &= 4 + 1 = 5 < 6 = D^4_{4,1} \end{aligned}$$

et $D^4_{i,5} + D^4_{5,j} \geqslant D^4_{i,j}$ dans les autres cas, d'où

$$D^5 = \begin{bmatrix} 0 & 2 & 1 & 5 & 4 \\ 3 & 0 & 3 & 6 & 2 \\ 1 & 1 & 0 & 4 & 3 \\ 5 & 2 & 5 & 0 & 4 \\ 1 & 1 & 2 & 6 & 0 \end{bmatrix} \qquad P^5 = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 \\ 5 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 2 \\ 2 & 2 & 2 & 4 & 2 \\ 1 & 2 & 1 & 1 & 5 \end{bmatrix}.$$

Dans le tableau D^5, le nombre $D^5_{i,j}$ est la durée d'un plus court chemin de i vers j. Pour trouver un tel chemin, on utilise le tableau P^5, car $P^5_{i,j}$ est le sommet qui suit i dans un plus court chemin de i vers j.
Cherchons par exemple un plus court chemin de 4 vers 3 : on a $P^5_{4,3} = 2$ donc le chemin commence par les sommets $4, 2$; le sommet suivant est $P^5_{2,3} = 3$, donc $(4, 2, 3)$ est un plus court chemin pour aller de 4 vers 3 ; sa durée est $D^5_{4,3} = 5$.

De même, pour trouver un plus court chemin allant de 1 à 5, on observe que $P^5_{1,5} = 3$, donc le chemin commence par $1, 3$; les sommets suivants sont $P^5_{3,5} = 2$, puis $P^5_{2,5} = 5$; il s'agit donc du chemin $(1, 3, 2, 5)$ de poids $D^5_{1,5} = 4$.

Algorithme de recherche d'un chemin de poids minimum

On numérote de 1 à n les sommets du graphe et l'on suppose que pour chaque arc $\widehat{i, j}$, on a défini un poids $d_{i,j}$ pour le chemin (i, j) et un poids $d_{j,i}$ pour le chemin (j, i). S'il n'y a pas d'arc entre les sommets i et j, on définit ces nombres en leur donnant une valeur très grande par rapport aux poids.

Étape 1. On construit les tableaux D et P à n lignes et n colonnes en posant $D_{i,j} = d_{i,j}$ et $P_{i,j} = j$ pour $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant n$ (i est l'indice de la ligne, j celui de la colonne).

Étape 2. Pour k de 1 à n, exécuter la tâche suivante :

pour i de 1 à n, pour j de 1 à n,
si $D_{i,k} + D_{k,j} < D_{i,j}$, alors $\big(D_{i,j} \leftarrow D_{i,k} + D_{k,j}$ et $P_{i,j} \leftarrow P_{i,k}\big)$.

Après exécution de cet algorithme, $D_{i,j}$ est le poids minimum d'un chemin de i vers j. Les sommets successifs d'un tel chemin sont $i, i_1 = P_{i,j}, i_2 = P_{i_1,j}$, etc. Mis à part le point de départ i, ces sommets se lisent en parcourant dans un ordre convenable la j-ième colonne de P.

3.3 Le problème du flot maximum

Exemple. À partir d'une station de pompage située en s, un réseau de conduites permet d'acheminer du pétrole en des lieux a, b, c, d et t. Ci-dessous le plan du réseau. La flèche indique le sens d'écoulement dans chaque conduite et le chiffre entre parenthèses indique la capacité de la conduite, c'est-à-dire le débit maximum, en barils par heure. On obtient un graphe dont les arcs représentent les conduites. Mais chaque arc possède une origine et une extrémité, donc nous désignons les arcs par des couples : (s, a), (s, c), (c, a), (a, b), (c, d), (c, b), (b, d), (b, t) et (d, t).
Le graphe est dit *orienté*.

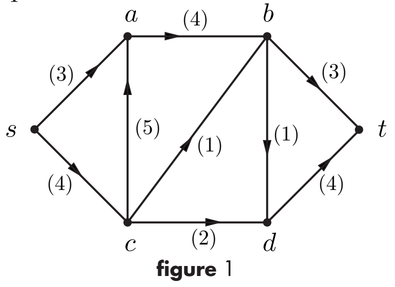


figure 1

On veut faire transiter de s à t le plus de pétrole possible en réglant convenablement le débit dans chaque conduite. Un régime d'écoulement est déterminé par

- le débit F assuré par la station de pompage s,
- la capacité de chaque conduite (u, v)
- la quantité de pétrole $f(u, v)$ qui transite en une heure dans la conduite (u, v).

La quantité pompée en s étant d'abord envoyée vers a et c, on a

$$f(s, a) + f(s, c) = F$$

En un point donné du réseau, il y a toujours autant de pétrole qui arrive que de pétrole qui part : cette loi de conservation se traduit par

$$f(s,a)+f(c,a)=f(a,b)\,,\qquad f(s,c)=f(c,a)+f(c,b)+f(c,d)$$
$$f(a,b)+f(c,b)=f(b,d)+f(b,t)\,,\qquad f(d,t)=f(c,d)+f(b,d)$$

et puisque tout le pétrole arrive en t, on aura aussi

$$f(b,t)+f(d,t)=F$$

Il s'agit de trouver les flux $f(u,v)$ pour que F soit maximum, sachant que, dans chaque conduite, le flux ne peut excéder la capacité. Voici la formulation du problème :

Trouver les $f(u,v)$ rendant F maximum, sachant que l'on a

$$\begin{aligned} f(s,a)+f(s,c)=f(b,t)+f(d,t) &= F\\ f(s,a)+f(c,a)-f(a,b) &= 0\\ f(s,c)-f(c,a)-f(c,b)-f(c,d) &= 0\\ f(a,b)+f(c,b)-f(b,d)-f(b,t) &= 0\\ f(c,d)+f(b,d)-f(d,t) &= 0 \end{aligned}$$

$$\begin{array}{c} 0\leqslant f(s,a)\leqslant 3\ ,\ 0\leqslant f(a,b)\leqslant 4\ ,\ 0\leqslant f(b,t)\leqslant 3\\ 0\leqslant f(s,c)\leqslant 4\ ,\ 0\leqslant f(c,d)\leqslant 2\ ,\ 0\leqslant f(d,t)\leqslant 4\\ 0\leqslant f(c,a)\leqslant 5\ ,\ 0\leqslant f(c,b)\leqslant 1\ ,\ 0\leqslant f(b,d)\leqslant 1 \end{array}$$

Énoncé du problème général

Soit G un graphe orienté et connexe.

- On suppose qu'il existe un unique sommet s, appelé *source*, où n'arrive aucun arc et qu'il existe un unique sommet t, appelé *terminal*, d'où ne part aucun arc.
- À chaque arc (u,v) est associé un nombre $c(u,v)\geqslant 0$, appelé *capacité* de l'arc ; s'il n'y a pas d'arc entre deux sommets u et v, on pose $c(u,v)=0$.
- Un *flot* est la donnée pour chaque arc (u,v) d'un nombre $f(u,v)\geqslant 0$ satisfaisant les conditions :
  i) $f(u,v)\leqslant c(u,v)$ pour tout arc (u,v).
  ii) Si a est un sommet différent de s et de t, alors $\sum_{\text{arc }(v,a)} f(v,a)=\sum_{\text{arc }(a,u)} f(a,u)$.
  La condition (ii) exprime la loi de conservation : en tout sommet différent de la source et du terminal, le flux entrant est égal au flux sortant.
- La *valeur* d'un flot f est la quantité $\text{val}(f)=\sum_{\text{arc }(s,u)} f(s,u)$, somme des flux issus de la source. Il résulte de la loi de conservation que l'on a aussi $\text{val}(f)=\sum_{\text{arc }(v,t)} f(v,t)$.

La valeur d'un flot ne peut pas excéder la somme des capacités des arcs aboutissant au terminal, ni celle des arcs issus de la source.

Problème : *trouver sur G un flot dont la valeur est la plus grande possible.*
Un tel flot s'appelle un *flot maximum*.

La recherche d'un flot maximum se rencontre dans la plupart des questions relatives aux réseaux de distribution, notamment en télécommunication.

Exemple. Reprenons le graphe de l'exemple présenté en introduction. Sur la figure suivante, nous indiquons à coté de chaque arc un couple capacité, flux : pour l'arc (s, a), on a ainsi $c(s, a) = 3$ et $f(s, a) = 2$. C'est un exemple de flot ; sa valeur est $\text{val}(f) = f(s, a) + f(s, b) = 4 = f(b, t) + f(d, t)$.

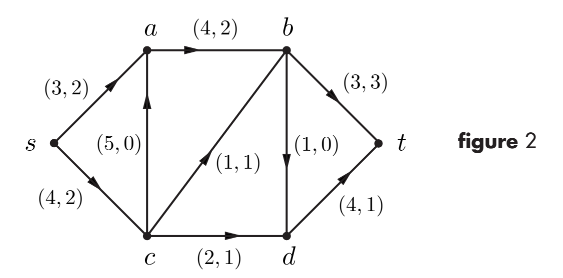


figure 2

Procédé d'augmentation d'un flot

Supposons que f est un flot (par exemple $f(u, v) = 0$ pour tout arc (u, v)).
Rappelons qu'un chemin de G est une suite de sommets adjacents deux à deux différents. Dans un chemin de G, il y a en général des arcs parcourus dans le sens de leur orientation et des arcs parcourus dans l'autre sens ; les premiers sont dits *directs*, les autres sont *indirects*. Par exemple, dans le cas du graphe de la figure 1, le chemin (s, a, b, c, d, t) a pour arcs directs (s, a), (a, b), (c, d) et (d, t) et l'arc (c, b) est indirect.
Supposons que γ est un chemin de s vers t dont tous les arcs sont directs et vérifient $f(u, v) < c(u, v)$. Si p est le minimum des différences $c(u, v) - f(u, v)$ le long de γ, on peut augmenter le flot de la valeur p sur tous les arcs du chemin, car la loi de conservation sera encore satisfaite après cette opération.
Par exemple, sur la figure 2, on peut augmenter le flot de 1 le long du chemin (s, c, a, b, d, t) ; le résultat est montré figure 3.

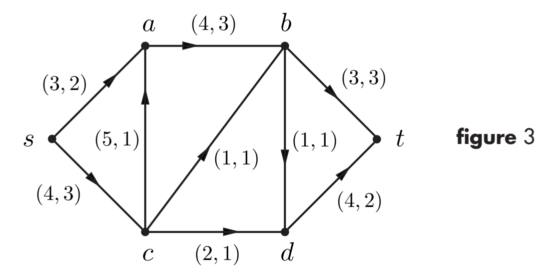


figure 3

Plus généralement, nous dirons qu'un chemin est *non saturé* si l'on a $f(u, v) < c(u, v)$ pour tous les arcs directs et $f(u, v) > 0$ pour tous les arcs indirects.
Supposons que γ est un chemin non saturé de s vers t. Pour tout arc α de γ, on pose

$$p(\alpha) = \begin{cases} c(\alpha) - f(\alpha) & \text{si } \alpha \text{ est direct} \\ f(\alpha) & \text{si } \alpha \text{ est indirect.} \end{cases}$$

On définit

$$p(\gamma) = \min_{\text{arc } \alpha \text{ de } \gamma} \{p(\alpha)\}$$

qui est un nombre strictement positif puisque γ est non saturé. Pour chaque arc du graphe, on pose

$$\hat{f}(\alpha) = f(\alpha) + p(\gamma) \quad \text{si } \alpha \text{ est un arc direct de } \gamma$$
$$\hat{f}(\alpha) = f(\alpha) - p(\gamma) \quad \text{si } \alpha \text{ est un arc indirect de } \gamma$$
$$\hat{f}(\alpha) = f(\alpha) \quad \text{si } \alpha \text{ n'est pas un arc de } \gamma.$$

On vérifie facilement que $\hat{f}$ est encore un flot et que l'on a $\text{val}(\hat{f}) = \text{val}(f) + p(\gamma)$. Puisque $p(\gamma)$ est strictement positif, on a $\text{val}(\hat{f}) > \text{val}(f)$.
Pour le flot f de la figure 3, le chemin $\gamma = (s,a,b,c,d,t)$ est non saturé : en effet, sur les arcs directs, le flot est strictement inférieur à la capacité et sur l'arc indirect (c,b), le flot est strictement positif. On a $p(s,a) = p(a,b) = p(c,b) = p(c,d) = 1$, $p(d,t) = 2$, donc $p(\gamma) = 1$. On peut donc augmenter le flot de 1 sur les arcs directs (s,a), (a,b), (c,d) et (d,t) et le diminuer de 1 sur (c,b) : on obtient le flot $\hat{f}$ de la figure 4.

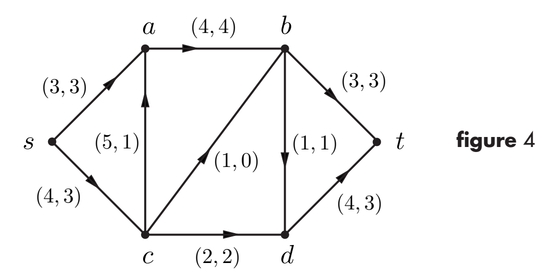


figure 4

Recherche de chemins non saturés

Pour chercher des chemins non saturés, on utilise un procédé de marquage successif des sommets du graphe. Voici la règle du marquage.

- La source s reçoit pour marque $(-,\infty)$, où le signe ∞ représente un nombre très grand par rapport aux capacités des arcs du graphe.
- Supposons que le sommet u a été marqué et que v est un sommet non marqué adjacent à u.

marquage en avant : si l'arc orienté est (u,v) et si $f(u,v) < c(u,v)$, on donne à v la marque (u^+, p_v), où $p_v = \min\{p_u, c(u,v) - f(u,v)\}$.

marquage en arrière : si l'arc orienté est (v,u) et si $f(v,u) > 0$, on donne à v la marque (u^-, p_v), où $p_v = \min\{p_u, f(v,u)\}$.

Quand un sommet v a été marqué, on peut remonter jusqu'à la source les prédécesseurs de v dans le marquage : en effet, si la marque de v contient u^+ ou u^-, c'est que le marquage de u a immédiatement précédé celui de v. Cela détermine un chemin $(s = a_0, a_1, a_2, \ldots, a_k = v)$ où a_{i+1} a été marqué à partir de a_i. Par construction, la marque p_v est strictement positive. Si $v{=}t$, on a obtenu un chemin non saturé de s à t.

Méthode de recherche d'un flot maximum

Dans une première phase, on marque les sommets successivement en partant de la source s. Cette phase se termine lorsque le terminal t reçoit une marque ou bien lorsqu'aucun sommet ne peut plus recevoir de marque.
Si t a été marqué, on a trouvé un chemin γ non saturé de s à t :

- on augmente alors le flot de la quantité $p(\gamma)$, minimum des marques p_v, où v parcourt les sommets de γ : on a donc simplement $p(\gamma) = p_t$.
- On supprime ensuite toutes les marques sauf celle de s, puis on recommence le marquage.
- S'il n'y a plus de sommet à marquer et si t n'a pas été marqué, le flot est maximum, du moins dans le cas où les capacités ont des valeurs entières : dans ce cas en effet, le flot augmente, à partir du flot nul, par valeurs entières, donc la valeur des flots successifs finit par être stationnaire.

Algorithme du flot maximum. On suppose que les capacités sont des entiers positifs ou nuls.

I) Trouver un flot dans le graphe G (on peut prendre $f(\alpha) = 0$ pour tout arc α) et donner à s la marque $(-, \infty)$.

II) Marquer les sommets successivement jusqu'à ce que t soit marqué ou bien qu'il n'y ait plus de sommet marquable. Si t est marqué, aller en (III) ; sinon aller en (V).

III) Poser $v = t$ et $p = p_t$. Tant que $v \neq s$ faire :

   i) si la marque de v contient u^+, alors $f(u,v) \leftarrow f(u,v) + p$;

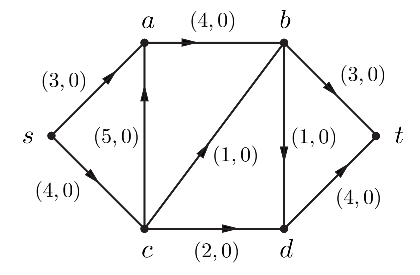
   ii) si la marque de v contient u^-, alors $f(v,u) \leftarrow f(v,u) - p$;

   iii) $v \leftarrow u$.

IV) Supprimer toutes les marques sauf celle de s et aller en (II).

V) Fin de l'algorithme : le flot est maximum.

Exemple. Appliquons l'algorithme à l'exemple précédent. Le graphe est celui de la figure 1 reproduite ci-dessous



On marque la source s par $(-, \infty)$ et l'on part du flot nul.

a) On peut marquer successivement les sommets c, a, b et t (phase II) :

► marquage de c à partir de s : la capacité de (s, c) est 4 et $f(s, c) = 0$, donc $p_c = \min\{\infty, 4 - 0\} = 4$ et c reçoit la marque $(s^+, 4)$;
► marquage de a à partir de c : la capacité de (c, a) est 5, $f(c, a) = 0$, donc $p_a = \min\{p_c, 5 - 0\} = 4$ et a reçoit la marque $(c^+, 4)$;
► marquage de b à partir de a : $p_b = \min\{p_a, c(a, b) - f(a, b)\} = 4$, donc b reçoit la marque $(a^+, 4)$;
► marquage de t à partir de b : on a $p_t = \min\{p_b, c(b, t) - f(b, t)\} = \min\{4, 3\} = 3$ donc t reçoit la marque $(b^+, 3)$.

Augmentons le flot de la quantité $p_t = 3$ sur le chemin (s, c, a, b, t) : il vient $f(b,t) = f(a,b) = f(c,a) = f(s,c) = 3$ (phase III) et supprimons toutes les marques sauf celle de s. On a obtenu le flot de la figure 5.

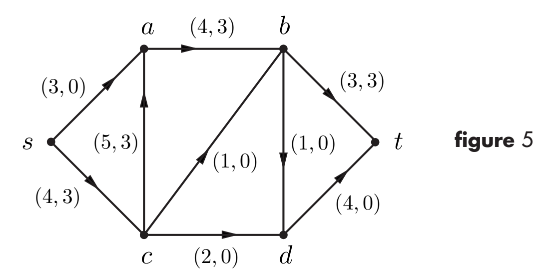


figure 5

b) Marquons successivement les sommets c, b, d, t.

► marquage de c : la capacité de (s, c) est 4, donc $p_c = \min\{\infty, 4\} = 4$ et c est marqué $(s^+, 4)$;
► marquage de b à partir de c : la capacité de (c,b) est 1, donc $p_b = \min\{p_c, 1\} = 1$ et b est marqué $(c^+, 1)$;
► marquage de d à partir de b : la capacité de (b,d) est 1, donc $p_d = \min\{p_b, 1\} = 1$, et d est marqué $(b^+, 1)$;
► marquage de t à partir de d : la capacité de (d,t) est 4, donc $p_t = \min\{p_d, 4\} = 1$ et t est marqué $(d^+, 1)$.

On augmente le flot de 1 sur le chemin (s, c, b, d, t), en posant $f(d, t) = f(b, d) = f(c, b) = 1$ et $f(s, c) = 3 + 1 = 4$ puis on supprime toutes les marques sauf celle de s. Le flot obtenu est montré figure 6.

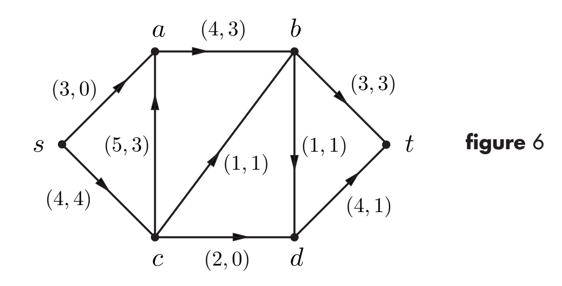


figure 6

c) Sur l'arc (s, c), le flot est égal à la capacité, donc on ne peut pas marquer c à partir de s.

- Le sommet a est marquable : $p_a = c(s, a) - f(s, a) = 3$, donc a reçoit la marque $(s^+, 3)$. On peut alors marquer b, mais ensuite ni d ni t ne sont marquables, car sur les arcs (b, d) et (b, t), le flot est égal à la capacité.
- À partir de a, on peut marquer c (marquage en arrière), car le flot de c vers a est positif. On obtient $p_c = \min\{p_a, 3\} = 3$ et c reçoit la marque $(a^-, 3)$.
- On peut maintenant marquer successivement d et t : le sommet d est marqué $(c^+, 2)$ et pour t, on a $p_t = \min\{p_d, c(d, t) - f(d, t)\} = \min\{2, 4 - 1\} = 2$, donc la marque de t est $(d^+, 2)$.

Le chemin parcouru au cours du marquage est (s, a, c, d, t) ; sur les arcs directs (d, t), (c, d) et (s, a), on augmente le flot de $p_t = 2$ et sur l'arc indirect (c, a), on diminue le flot de 2. On obtient le flot figure 7.

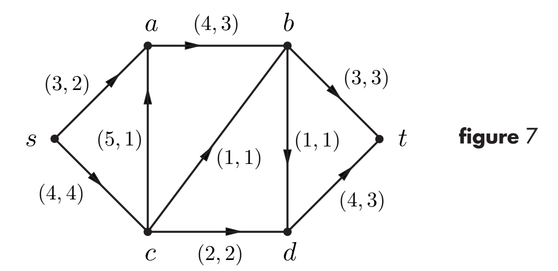


figure 7

d) Sur l'arc (s, c), le flot est égal à la capacité, donc on ne peut pas marquer c à partir de s. Le sommet a peut être marqué, car $c(s, a) - f(c, a) = 3 - 2 = 1 > 0$. On peut aussi marquer b et c à partir de a. Mais sur les arcs (b, t), (b, d) et (c, d), le flot est égal à la capacité : il sera donc impossible de marquer t. Les phases II, III, IV de l'algorithme sont exécutées, donc le flot obtenu est maximum (figure 7). La valeur du flot maximum est $\text{val}(f) = f(s, a) + f(s, c) = 2 + 4 = 6$.

En faisant d'autres choix pour les marquages, on peut obtenir d'autres flots maximum de même valeur, comme ceux de la figure 8.

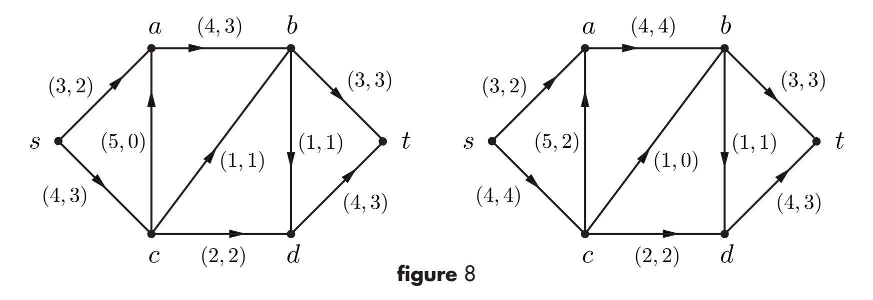


figure 8

Flot maximum et coupure minimum

Revenons au problème général et formulons quelques définitions.
Une *coupure* de G est une partition de l'ensemble des sommets en deux parties S, S' telles que $s \in S$ et $t \in S'$; les parties S et S' sont donc disjointes et leur réunion est l'ensemble de tous les sommets.

► La *capacité* d'une coupure (S, S') est le nombre $C(S,S') = \sum_{\substack{\text{arc } (u,u') \\ u \in S \text{ , } u' \in S'}} c(u,u')$,

► son *flot* est le nombre $F(S,S') = \sum_{\substack{\text{arc } (u,u') \\ u \in S \text{ , } u' \in S'}} f(u,u')$.

On définit de même le nombre $F(S', S)$.

Exemple. Pour l'un des flots représentés figure 8, prenons comme coupure $S = \{s,a,b,c\}$ et $S' = \{d,t\}$. Les arcs orientés ayant leur origine dans S et leur extrémité dans S' sont (c,d), (b,d) et (b,t), donc la capacité est $C(S,S') = 2+1+3 = 6$; le flot de cette coupure est aussi $F(S,S') = 2+1+3 = 6$ et l'on a $F(S',S) = 0$, car il n'y a aucun arc d'origine d ou t ayant son extrémité dans S.

Proposition. *Pour tout flot f et pour toute coupure (S,S'), on a*

$$\text{val}(f) = F(S,S') - F(S',S)\,.$$

Démonstration. Pour tout sommet $u \in S$, on a

$$\sum_v f(u,v) - \sum_v f(v,u) = \begin{cases} \text{val}(f) & \text{si } u = s \\ 0 & \text{si } u \in S \setminus \{s\} \end{cases}$$

car f est un flot. En ajoutant ces égalités pour les différents sommets de S, il vient

$$\sum_{u \in S} \sum_v f(u,v) - \sum_{u \in S} \sum_v f(v,u) = \text{val}(f)\,.$$

Si u et v sont deux sommets de S, les termes $f(u,v)$ et $-f(u,v)$ apparaissent exactement une fois dans chaque somme, donc se détruisent. Après simplification, on obtient l'égalité

$$\sum_{u \in S} \sum_{v \in S'} f(u,v) - \sum_{u \in S} \sum_{v \in S'} f(v,u) = \text{val}(f)\,,$$

c'est-à-dire $F(S,S') - F(S',S) = \text{val}(f)$. ■

Corollaire. *Soient f un flot et (S,S') une coupure de G.*

i) *On a* $\text{val}(f) \leqslant F(S,S')$.

ii) *Si* $\text{val}(f) = C(S,S')$, *alors le flot f est maximum et (S,S') est une coupure de capacité minimum.*

Démonstration. On a $\text{val}(f) = F(S,S') - F(S',S) \leqslant F(S,S')$, car $F(S',S) \geqslant 0$. Soit f^* un flot maximum et (S^*, S'^*) une coupure de capacité minimum. D'après (i), on a

$$\text{val}(f^*) \leqslant F(S^*,S'^*) \leqslant C(S^*,S'^*)\,, \text{ d'où}$$
$$\text{val}(f) \leqslant \text{val}(f^*) \leqslant C(S^*,S'^*) \leqslant C(S,S')\,.$$

Supposons $\text{val}(f) = C(S,S')$. Il s'ensuit $\text{val}(f) = \text{val}(f^*) = C(S^*,S'^*) = C(S,S')$, donc f est un flot maximum et (S,S') est une coupure de capacité minimum. ■

Pour justifier l'algorithme, il reste à montrer que si un flot à capacités entières ne présente pas de chemin non saturé de s à t, alors ce flot est maximum.

Proposition. *Un flot à capacités entières est maximum si et seulement s'il n'y a pas de chemin non saturé de s à t.*

Démonstration. S'il y a un chemin non saturé de s à t pour le flot f, le procédé d'augmentation conduit à un flot f^* de valeur strictement supérieure : le flot f n'est donc pas maximum. Réciproquement, supposons que pour le flot f, il n'y a pas de chemin non saturé de s à t. Soit S l'ensemble des sommets qu'on peut atteindre à partir de s par un chemin non saturé. On a $s \in S$, et par hypothèse $t \notin S$. En notant S' l'ensemble des sommets qui ne sont pas dans S, on définit donc une coupure (S, S'). Soit (u, v) un arc tel que $u \in S$ et $v \in S'$. Par définition de S, il existe un chemin γ non saturé de s à u. Si l'on ajoute l'arc (u, v) à ce chemin, on obtient un chemin saturé, car $v \notin S$. On a donc $f(u, v) = c(u, v)$. De même, si (w, z) est un arc tel que $w \in S'$ et $z \in S$, on a $f(w, z) = 0$. Il s'ensuit $F(S, S') = C(S, S')$ et $F(S', S) = 0$. Par suite $\mathrm{val}(f) = F(S, S') - F(S', S) = C(S, S')$. D'après le corollaire précédent, f est un flot maximum et (S, S') est une coupure de capacité minimum. ■

Puisque l'algorithme se termine lorsqu'il n'y a plus de chemin non saturé de s à t, on en déduit qu'il fournit un flot maximum. D'après la seconde partie de la démonstration ci-dessus, on a le théorème suivant.

Théorème du flot maximum. *Si les capacités sont entières, la valeur du flot maximum est égale à la plus petite des capacités des coupures.*

Remarques

- On a supposé que les capacités sont des entiers positifs ou nuls afin d'assurer que l'algorithme se termine. Cette restriction est sans importance dans les applications, car en choisissant des unités convenables pour les capacités, on peut toujours se ramener à ce cas.
- La longueur de l'algorithme est conditionnée par la taille des capacités : le temps d'exécution peut donc être grand, même pour un graphe ayant peu d'arcs (exercice 14). Pour augmenter sensiblement l'efficacité de l'algorithme, il faut, dans la phase de marquage, chercher systématiquement un chemin non saturé de s à t ayant le moins d'arcs possible.

Généralisations

Voici deux généralisations au problème du flot maximum.

Graphe ayant plusieurs sources ou plusieurs terminaux

On conserve les hypothèses générales faites sur G page 87, mais on ne suppose plus l'unicité de la source et du terminal. Au contraire, il peut y avoir des sommets $s_1, \ldots, s_p$ où n'arrive aucun arc (ce sont les sources) et des sommets $t_1, \ldots, t_q$ d'où ne part aucun arc (ce sont les terminaux). On définit un flot de la même façon, la loi de conservation devant être vérifiée en chaque sommet différent d'une source ou d'un terminal.

La valeur d'un flot f est le nombre $\text{val}(f)=\sum_{1\leqslant i\leqslant p}\text{val}_i(f)$, où $\text{val}_i(f)=\sum_{\text{arc}\,(s_i,u)} f(s_i,u)$ est la somme des flux issus de la source s_i.

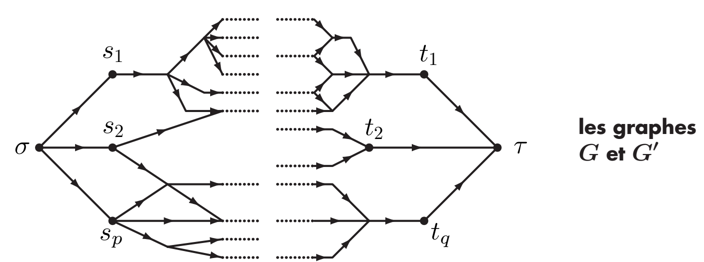
On a aussi $\text{val}(f) = \sum_{1\leqslant j\leqslant q}\ \sum_{\text{arc}\,(v,t_j)} f(v,t_j)$, d'après la loi de conservation.

Pour trouver un flot maximum sur G, nous allons nous ramener au cas d'une seule source et d'un seul terminal. Pour cela, on définit un graphe orienté G' de la manière suivante :

- on ajoute à G deux sommets σ et τ,
- pour chaque source s_i, on ajoute un arc (σ, s_i) et pour chaque terminal t_j, on ajoute un arc (t_j, τ).

Le graphe G' a pour unique source σ et pour unique terminal τ. Définissons des capacités très grandes sur les arcs qui ont été ajoutés et gardons les capacités sur les arcs de G. Tout flot sur G' définit un flot sur G ; réciproquement, si f est un flot sur G, on définit un flot f' sur G' en posant $f'(\alpha) = f(\alpha)$ si α est un arc de G, $f'(\sigma, s_i) = \text{val}_i(f)$ et $f'(t_j, \tau) = \sum_{\text{arc}\,(v,t_j)} f(v, t_j)$.

Les flots f et f' ayant même valeur, il suffit d'appliquer à G' l'algorithme de recherche d'un flot maximum : on obtiendra ainsi un flot maximum sur G.



les graphes G et G'

Flot à double contraintes

Dans les applications, on doit parfois munir chaque arc α d'une capacité minimum $m(\alpha) \geqslant 0$ et d'une capacité maximum $M(\alpha)$. Un flot f est dit *admissible* si l'on a $m(\alpha) \leqslant f(\alpha) \leqslant M(\alpha)$ pour tout arc α. Si tous les $m(\alpha)$ sont nuls, le flot nul est admissible et l'algorithme fournit un flot admissible maximum. Mais si certains nombres $m(\alpha)$ sont strictement positifs, le flot nul n'est pas admissible.

- Si l'on a trouvé un flot f admissible, alors on peut pratiquer l'algorithme du flot maximum en partant de ce flot f. On obtient ainsi un flot admissible maximum.
- En général, la question se pose de trouver un flot admissible ; un tel flot n'existe d'ailleurs pas toujours. Dans le graphe ci-dessous, nous avons indiqué à coté de chaque arc, l'intervalle $[m, M]$ pour un flot admissible : on vérifie facilement qu'il n'existe pas de flot admissible.

Dans l'exercice 16, on présente une méthode pour chercher un flot admissible.

Exercices

@ **1.** Un photocopieur est mis à la disposition de n personnes ($n \leqslant 200$). Chaque utilisateur possède un code personnel à quatre chiffres permettant d'activer le photocopieur.

a) Combien y a-t-il de codes possibles ? Combien sont actifs ?

b) Le nombre d'essais pour rentrer un code est limité à trois. On rentre un code au hasard ; s'il n'est pas bon, on en rentre un deuxième et éventuellement un troisième. Quelle est la probabilité p pour qu'on active ainsi le photocopieur ? (on pourra chercher la probabilité $q = 1 - p$ pour que les trois essais soient infructueux).

c) Posons $x = 10^{-4}n$. Montrer que l'on a $3x(1-x) \leqslant p \leqslant \dfrac{3x}{1 - 2\,10^{-4}}$. On veut que la probabilité p reste inférieure à 5% : estimer le nombre d'utilisateurs à ne pas dépasser.

2. Un événement se produit avec la probabilité 0,25. On veut réaliser une suite d'épreuves indépendantes où l'on a neuf chances sur dix que l'événement se produise entre vingt et trente fois sur cent. Combien d'épreuves suffit-il de faire ?

@ **3. Estimation asymptotique d'un indice de dispersion.** Pour quantifier la dispersion d'une variable aléatoire X, il est naturel de prendre l'espérance d de la variable aléatoire $|X - E|$ (bien qu'on préfère plutôt choisir l'écart-type qui se comporte mieux dans les calculs). Supposons que X est le nombre de fois qu'un événement de probabilité p se produit au cours d'une suite de n épreuves indépendantes, autrement dit, X suit une loi binomiale : la probabilité que l'événement survienne k fois en n épreuves est $P(k) = \binom{n}{k}p^k(1-p)^{n-k}$. On suppose $n = 2N$ et $p = 1/2$.

a) Montrer que l'on a $d = \sum_{k=0}^{n}|k - N|P(k)$ et que l'écart-type de X est $\sigma = \sqrt{N/2}$.

b) En admettant l'égalité $\sum_{k=0}^{N}(N-k)\binom{2N}{k} = \frac{N}{2}\binom{2N}{N}$, démontrer que $d = \dfrac{N}{2^{2N}}\dfrac{(2N)!}{(N!)^2}$.

c) Quand N tend vers l'infini, on a $N! \sim N^N e^{-N}\sqrt{2\pi N}$ (formule de Stirling). Montrer que d/σ tend vers $\sqrt{2/\pi}$ quand N tend vers l'infini.

4. On repère les points de l'espace par des coordonnées cartésiennes. Les points à coordonnées entières sont les sommets de parallélépipèdes à arêtes parallèles aux axes. Comme pour un quadrillage dans le plan, considérons les chemins formés d'arêtes orientées dans le sens des coordonnées croissantes. Si p, q, r sont des entiers positifs, montrer qu'il y a $\dfrac{(p+q+r)!}{p!\,q!\,r!}$ chemins joignant l'origine au point de coordonnées (p, q, r).

@ **5.** Soient n, a, b des entiers positifs ou nuls. On cherche le nombre N de solutions (x,y,z) de l'équation $x + y + z = n$, où x, y, z sont des entiers positifs vérifiant $1 \leqslant x \leqslant a$ et $1 \leqslant y \leqslant b$. On pose $S = \{(x,y,z) \mid x, y, z \text{ entiers}, \ x \geqslant 1, \ y \geqslant 1, \ z \geqslant 1, \ x + y + z = n\}$,

$$A = \{(x,y,z) \in S \mid x > a\}, \quad B = \{(x,y,z) \in S \mid y > b\}.$$

a) Montrer que N est le nombre d'éléments de l'ensemble $A' \cap B'$, où $A' = S \setminus A$ et $B' = S \setminus B$.

b) Montrer que $|A' \cap B'| = |S| - |A| - |B| + |A \cap B|$.

c) Montrer que $|A|$ est le nombre de solutions (u, y, z) de l'équation $u + y + z = n - a$, où u, y, z sont des entiers strictement positifs (s'inspirer de l'exemple page 64). En déduire $|A| = \binom{n-a-1}{2}$.

d) Montrer de même que l'on a $|B| = \binom{n-b-1}{2}$ et $|A \cap B| = \binom{n-a-b-1}{2}$. En déduire la valeur de N.

6. Un vacancier veut envoyer des cartes postales à quatre personnes. Il achète sept cartes différentes. De combien de façons peut-il toutes les expédier ?

@ **7.** Pour tout entier $n \geqslant 1$, posons $E_n = \{1, 2, \ldots, n\}$.

a) Soit $f : E_{n+1} \to E_n$ une application. Montrer que si $f(E_{n+1}) = E_n$, il existe un unique élément $b \in E_n$ ayant deux antécédents.

b) En déduire qu'il y a $n! \dfrac{n(n+1)}{2}$ applications de E_{n+1} dans E_n ayant pour image E_n.

8. Autre calcul du nombre d'applications ayant pour image leur ensemble d'arrivée (page 66). Soient p et n des entiers positifs tels que $p \geqslant n$. Si U est un ensemble à p éléments et si V est un ensemble à n éléments, on note $s(p, n)$ le nombre d'applications $h : U \to V$ telles que $h(U) = V$. On pose $E_p = \{1, 2, \ldots, p\}$ et $E_n = \{1, 2, \ldots, n\}$.

a) Soit B une partie de E_n. Montrer que si B possède k éléments, il y a $s(p, k)$ applications $f : E_p \to E_n$ telles que $f(E_p) = B$.

b) En déduire l'égalité $n^p = \sum_{k=1}^{n} \binom{n}{k} s(p, k)$.

c) Expliquer comment cette égalité permet de calculer les nombres $s(p, k)$ de proche en proche.

9. Nombre de partitions d'un ensemble. Soit E un ensemble à p éléments et soit n un entier tel que $1 \leqslant n \leqslant p$. Une n-partition de E est la donnée de n parties $A_1, A_2, \ldots, A_n$ de E formant une partition de E : cela signifie que les parties A_i sont non vides, deux à deux disjointes et que leur réunion est E.

a) Expliquer comment une application $f : E \to \{1, 2, \ldots, n\}$ telle que $f(E) = \{1, 2, \ldots, n\}$ détermine une n-partition de E : considérer les parties $A_i = \{x \in E \mid f(x) = i\}$.

b) Montrer qu'il y a autant de n-partitions de E que d'applications $f : E \to \{1, 2, \ldots, n\}$ telles que $f(E) = \{1, 2, \ldots, n\}$.

@ **10. Étude d'une permutation**

a) Décomposer en cycles à supports disjoints la permutation s suivante :

i	1	2	3	4	5	6	7	8	9	10	11	12
$s(i)$	9	12	8	1	4	3	2	10	5	11	6	7

b) Quel est le plus petit entier $r \geqslant 1$ tel que s^r soit l'application identité ?

c) Calculer explicitement la bijection s^{100} (faire la division euclidienne de 100 par r).

@ 11. Calculs dans le groupe des permutations de six objets

a) Combien y a-t-il de permutations d'un ensemble à six éléments ?

b) Pour une permutation de l'ensemble $E = \{1, 2, 3, 4, 5, 6\}$, quelles sont les possibilités de décomposition en cycles à supports disjoints ? En déduire que si s est une permutation de l'ensemble E, alors s^{12} est égale à l'identité ou à s^2.

12. Une utilisation des graphes en Biologie. L'étude de l'évolution moléculaire conduit à définir des écarts entre certaines séquences d'ADN. Les distances entre ces séquences permettent de faire des hypothèses sur l'ordre dans lequel s'effectuent les duplications de gènes. Voici une table des distances entre six séquences impliquées dans la synthèse de protéines fixatrices d'oxygène.

	α	β	ε	γ	μ
β	5,6	0			
ε	6,1	2,4	0		
γ	5,9	2,7	2	0	
μ	7,5	7,6	7,7	7,6	0
ζ	4	6,4	6	6	8

a) Trouver un arbre de longueur minimum reliant ces séquences.

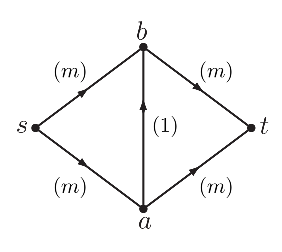
b) Calculer la longueur d'un plus court chemin entre deux séquences données.

13. Soit G un graphe orienté connexe ayant une source, un terminal et une capacité pour chaque arc. On suppose qu'il existe un cycle $C = (u_0, u_1, \ldots, u_{p-1}, u_0)$ que l'on peut parcourir en respectant l'orientation des arcs.

a) Montrer que les sommets s et t ne sont pas dans le cycle C.

b) Soit f un flot sur G et soit k un nombre inférieur aux valeurs $c(\alpha) - f(\alpha)$, où α sont les arcs du cycle C. Pour chaque arc α de G, posons $f'(\alpha) = f(\alpha)$ si α n'est pas un arc de C et $f'(\alpha) = f(\alpha) + k$ sinon. Montrer que f' est un flot sur G et que l'on a $\mathrm{val}(f') = \mathrm{val}(f)$.

14. Considérons le graphe orienté ci-contre, où la capacité de chaque arc est indiquée entre parenthèses. Le nombre m est un entier, s est la source et t est le terminal.

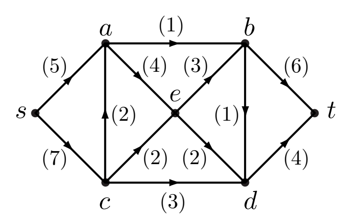


a) On part du flot nul et l'on pratique l'algorithme du flot maximum en utilisant, lors du marquage, alternativement les chemins $\gamma_1 = (s, a, b, t)$ et $\gamma_2 = (s, b, a, t)$. De combien augmente la valeur du flot à chaque étape ?

b) Montrer qu'il faut $2m$ augmentations du flot pour obtenir un flot maximum.

@ 15. Recherche d'un flot maximum. Pour le graphe orienté ci-dessous, les chiffres entre parenthèses indiquent la capacité des arcs. Trouver le flot maximum pouvant transiter

de la source s vers le terminal t.



16. Recherche d'un flot admissible dans le cas de doubles contraintes. Soit G un graphe orienté vérifiant les hypothèses faites page 87. Pour chaque arc α de G, on note $m(\alpha)$ la capacité minimum de l'arc et $M(\alpha)$ sa capacité maximum, en supposant $0 \leqslant m(\alpha) \leqslant M(\alpha)$. Soit G^* le graphe orienté obtenu de la manière suivante :

- on ajoute à G deux sommets s^* et t^* ;
- pour chaque arc $\alpha=(u,v)$ de G, on ajoute un arc $\alpha'=(s^*,v)$ et un arc $\alpha''=(u,t^*)$;
- on définit les capacités sur G^* en posant, pour tout arc $\alpha=(u,v)$ de G :

$$c(\alpha') = \sum_{z \in G} m(z,v) \quad , \quad c(\alpha'') = \sum_{z \in G} m(u,z) \quad \text{et} \quad c(\alpha) = M(\alpha) - m(\alpha) \,;$$

- on ajoute un arc (t,s) de « capacité infinie ».

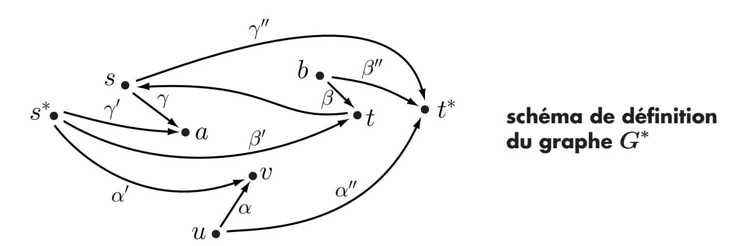


schéma de définition du graphe G^*

a) Vérifier que G^* est un graphe orienté ayant pour seule source s^* et pour seul terminal t^*.

Un flot f^* sur G^* est dit *saturant* si l'on a $f^*(\alpha') = c(\alpha')$ pour tout arc α de G.

b) Montrer qu'un flot saturant est un flot maximum sur G^*.

On suppose désormais que f^* est un flot saturant sur G^*. Pour tout arc α de G, on pose $f(\alpha) = f^*(\alpha) + m(\alpha)$.

c) Montrer que l'on a $m(\alpha) \leqslant f(\alpha) \leqslant M(\alpha)$ pour tout arc α de G.

d) Montrer l'égalité $\sum_{\text{arc}\,\alpha\,\text{de}\,G} f^*(\alpha') = \sum_{\text{arc}\,\alpha\,\text{de}\,G} f^*(\alpha) = \sum_{\text{arc}\,\alpha\,\text{de}\,G} m(\alpha)$.

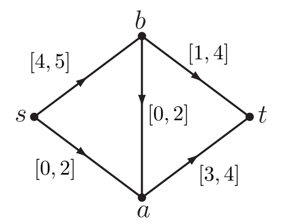
e) En déduire que pour tout arc α de G, on a $f^*(\alpha'') = c(\alpha'')$.

f) Soit u un sommet de G différent de s et de t. Montrer que l'on a

$$\textstyle\sum_{\text{arc}\,(y,u)\,\text{de}\,G} f(y,u) = \sum_{\text{arc}\,(u,z)\,\text{de}\,G} f(u,z)\,.$$

g) En déduire que f est un flot admissible sur G.

h) Pour le graphe G suivant, représenter le graphe G^*, trouver un flot saturant f^* sur G^* et calculer le flot admissible f. Le flot f est-il maximum ? Quelle est la valeur d'un flot maximum sur G ?



Voici une méthode pour rechercher un flot admissible sur G ou pour montrer qu'il n'en existe pas :

- trouver un flot maximum f^* sur G^*, en utilisant par exemple l'algorithme du flot maximum ;
- si f^* est saturant, alors le flot f défini comme ci-dessus est un flot admissible sur G ;
- si f^* n'est pas saturant, on peut montrer qu'il n'existe pas de flot admissible sur G.

Chapitre 4

Équations linéaires et vecteurs

1. Vecteurs et combinaisons linéaires

1.1 Les espaces vectoriels $\mathbb{R}^p$ et $\mathbb{C}^p$

Rappelons que $\mathbb{R}^p$ est l'ensemble des p-uplets $(x_1, x_2, \ldots, x_p)$ de nombres réels ; de même, $\mathbb{C}^p$ désigne l'ensemble des p-uplets de nombres complexes. Nous désignons par la lettre $\mathbb{K}$ l'ensemble $\mathbb{R}$ ou $\mathbb{C}$.

- Un élément de $\mathbb{K}^p$ s'appelle un *vecteur* à p coordonnées et, pour cette raison, l'ensemble $\mathbb{K}^p$ s'appelle un *espace vectoriel*.
- Le vecteur $(0, 0, \ldots, 0)$, dont toutes les coordonnées sont nulles, s'appelle le *vecteur nul* et se note simplement 0.
- Les éléments de $\mathbb{K}$ s'appellent des *scalaires* : ce sont des nombres.

Notation. Un vecteur $(x_1, \ldots, x_p)$ est souvent noté $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ et appelé alors une *matrice-colonne*.

Opérations sur les vecteurs de $\mathbb{K}^p$

Définitions

Soient $u = (x_1, x_2, \ldots, x_p)$ et $v = (y_1, y_2, \ldots, y_p)$ des vecteurs de $\mathbb{K}^p$.

- La somme des vecteurs u et v est le vecteur $u+v = (x_1+y_1, x_2+y_2, \ldots, x_p+y_p)$. Sous forme de matrices-colonne, cela s'écrit

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_p + y_p \end{bmatrix}.$$

- Si a est un nombre de $\mathbb{K}$, le produit du vecteur u par le scalaire a est le vecteur $au = (ax_1, ax_2, \ldots, ax_p)$, ou encore

$$a \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_p \end{bmatrix}.$$

- Soient $u_1, u_2, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$. Une *combinaison linéaire* des vecteurs $u_1, u_2, \ldots, u_n$ est un vecteur de la forme $a_1u_1 + a_2u_2 + \cdots + a_nu_n$, où $a_1, a_2, \ldots, a_n$ sont des scalaires.
- Si u est un vecteur non nul et a un scalaire, le vecteur au est dit *colinéaire à u.*

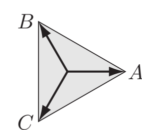
Si a, b sont des scalaires et u, v des vecteurs de $\mathbb{K}^p$, alors on a

$$au + bu = (a+b)u \quad \text{et} \quad a(u+v) = au + av.$$

Exemple 1. Posons $u = (2, 1, -1)$, $v = (3, 1, -2)$ et $w = (0, 1, 2)$. Si x, y, z sont des nombres réels, la combinaison linéaire $xu + yv + zw$ s'écrit matriciellement

$$x \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} + y \begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix} + z \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2x \\ x \\ -x \end{bmatrix} + \begin{bmatrix} 3y \\ y \\ -2y \end{bmatrix} + \begin{bmatrix} 0 \\ z \\ 2z \end{bmatrix} = \begin{bmatrix} 2x + 3y \\ x + y + z \\ -x - 2y + 2z \end{bmatrix}.$$

Exemple 2. Dans le plan euclidien $\mathbb{R}^2$, les points $A = (1, 0)$, $B = (-1/2, \sqrt{3}/2)$ et $C = (-1/2, -\sqrt{3}/2)$ sont les sommets d'un triangle équilatéral centré à l'origine O : la somme $\overrightarrow{OA} + \overrightarrow{OB} + \overrightarrow{OC}$ est le vecteur nul.



Définition

Pour tout indice i compris entre 1 et p, notons $\boldsymbol{e_i}$ le vecteur de $\mathbb{K}^p$ dont toutes les coordonnées sont nulles sauf celle d'indice i qui vaut 1 : on a

$$\boldsymbol{e_1} = (1, 0, 0, \ldots, 0)\ ,\ \boldsymbol{e_2} = (0, 1, 0, \ldots, 0)\ ,\ \ldots\ , \boldsymbol{e_p} = (0, 0, \ldots, 0, 1)$$

et matriciellement : $$\boldsymbol{E_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ \boldsymbol{E_2} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ \ldots, \boldsymbol{E_p} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Les vecteurs $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}$ s'appellent les *vecteurs canoniques*.

Pour tout vecteur $(x_1, \ldots, x_p) \in \mathbb{K}^p$, on a

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_p \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = x_1\boldsymbol{E_1} + x_2\boldsymbol{E_2} + \cdots + x_p\boldsymbol{E_p}.$$

Tout vecteur de $\mathbb{K}^p$ s'écrit donc de manière unique comme combinaison linéaire des vecteurs canoniques.

1.2 Sous-espaces vectoriels de $\mathbb{K}^p$

Définition

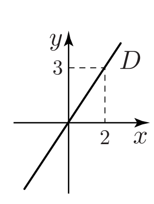
Soit V une partie de $\mathbb{K}^p$. On dit que V est un *sous-espace vectoriel* de $\mathbb{K}^p$ si

i) le vecteur nul appartient à V,

ii) pour tous vecteurs u et v appartenant à V, on a $u + v \in V$ et $au \in V$ quel que soit le scalaire $a \in \mathbb{K}$.

Exemples

► Soit D l'ensemble des $(x, y) \in \mathbb{R}^2$ tels que $3x - 2y = 0$.



Le vecteur nul $(0, 0)$ appartient à D. Supposons que les vecteurs $u{=}(x{,}y)$ et $u'{=}(x'{,}y')$ appartiennent à D, donc $3x{-}2y{=}0{=}3x'{-}2y'$. On a $3(x{+}x'){-}2(y{+}y'){=}(3x{-}2y){+}(3x'{-}2y'){=}0$, donc $u{+}u'{\in}D$. De même, pour tout $a \in \mathbb{R}$, on a $3(ax) - 2(ay) = a(3x - 2y) = 0$, donc $au \in D$.

L'ensemble D est donc un sous-espace vectoriel de $\mathbb{R}^2$. Géométriquement, D est la droite passant par l'origine et dirigée par le vecteur $(2, 3)$.

► Soient r un nombre réel et $P = \{(x, y, z) \in \mathbb{R}^3 \mid 3x - 2y + rz = 0\}$.

Supposons que les vecteurs $u = (x, y, z)$ et $u' = (x', y', z')$ appartiennent à P. On a $3(x{+}x'){-}2(y{+}y'){+}r(z{+}z'){=}(3x{-}2y{+}rz){+}(3x'{-}2y'{+}rz'){=}0$ et, pour tout $a{\in}\mathbb{R}$, $3(ax) - 2(ay) + r(az) = a(3x - 2y + rz) = 0$, donc $u + u'$ et au appartiennent à P.

Puisque le vecteur nul appartient à P, P est un sous-espace vectoriel de $\mathbb{R}^3$: c'est le plan passant par l'origine et orthogonal au vecteur $(3, -2, r)$; le plan P coupe le plan des x, y (d'équation $z = 0$) selon la droite D précédente.

► Soient $u_1, u_2, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$ et soit V l'ensemble de toutes les combinaisons linéaires des vecteurs $u_1, u_2, \ldots, u_n$: alors V est un sous-espace vectoriel de $\mathbb{K}^p$.

En effet, le vecteur nul est égal à la combinaison linéaire $0u_1 + 0u_2 + \cdots + 0u_n$, donc appartient à V. Soient $u{=}x_1u_1{+}\cdots{+}x_nu_n$ et $v{=}y_1u_1{+}\cdots{+}y_nu_n$ des vecteurs de V. On a $u + v = (x_1 + y_1)u_1 + \cdots + (x_n + y_n)u_n$, donc $u + v$ est un vecteur de V. De même, on a $au = (ax_1)u_1 + \cdots + (ax_n)u_n$, donc pour tout scalaire a, le vecteur au appartient à V.

Définition

Soient $u_1, u_2, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$. L'ensemble de toutes les combinaisons linéaires des vecteurs $u_1, u_2, \ldots, u_n$ est un sous-espace vectoriel de $\mathbb{K}^p$ appelé le *sous-espace vectoriel engendré par* $u_1, \ldots, u_n$.

1.3 Équations linéaires

On a souvent besoin de savoir si un vecteur $b{\in}\mathbb{K}^p$ est combinaison linéaire de vecteurs $u_1, u_2, \ldots, u_n$ donnés et si oui, de calculer les coefficients d'une telle combinaison.

Définition

Soient $u_1, u_2, \dots, u_n$ des vecteurs de $\mathbb{K}^p$ et soit $b \in \mathbb{K}^p$. Trouver tous les nombres $x_1, x_2, \dots, x_n$ tels que $x_1u_1 + x_2u_2 + \cdots + x_nu_n = b$ s'appelle *résoudre l'équation linéaire* $x_1u_1 + x_2u_2 + \cdots + x_nu_n = b$. Une solution de l'équation est donc un élément $(x_1, x_2, \dots, x_n) \in \mathbb{K}^n$.

Supposons que $u_1, u_2, \dots, u_n$ sont des vecteurs de $\mathbb{K}^p$. Pour désigner en général les coordonnées de ces vecteurs, on utilise des lettres munies de deux indices, en posant

$$u_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{pi} \end{bmatrix} \quad \text{pour } i = 1, 2, \dots, n\,.$$

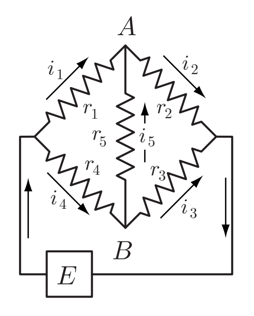
Le premier indice est le numéro de la coordonnée, le second est le numéro du vecteur. Pour tous scalaires $x_1, \dots, x_n$, la combinaison linéaire $x_1u_1 + x_2u_2 + \cdots + x_nu_n$ est égale à

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{pn} \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n \end{bmatrix}.$$

Si $b = (b_1, b_2, \dots, b_p)$, résoudre l'équation linéaire $x_1u_1 + x_2u_2 + \cdots + x_nu_n = b$ consiste donc à calculer les $(x_1, \dots, x_n) \in \mathbb{K}^n$ vérifiant les équations numériques

$$(S) \qquad \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\vdots \qquad\qquad\qquad\qquad \vdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = b_p \end{cases}$$

Un tel ensemble d'équations s'appelle un *système linéaire* de p équations à n inconnues. Les vecteurs $u_1, \dots, u_n$ sont les vecteurs-colonne du système.
Le vecteur b s'appelle le *second membre* du système.



Exemple. Considérons le circuit électrique ci-contre, du type « pont de Wheatstone ». Les résistances ont pour valeurs (en ohm) r_1, r_2, r_3, r_4, r_5 et E est un générateur. On a choisi un sens de parcours sur chaque portion du circuit, de sorte que les intensités i_1, i_2, i_3, i_4, i_5 sont des grandeurs algébriques. L'intensité dans le circuit principal est $i_1 + i_4 = i_2 + i_3$, en vertu de la loi des courants dérivés et l'on a de même $i_4 = i_3 + i_5$, $i_2 = i_1 + i_5$. D'après la loi d'Ohm, la différence de potentiel entre les points A et B est $r_5i_5 = r_1i_1 - r_4i_4 = r_3i_3 - r_2i_2$ et la différence de potentiel aux bornes du générateur est $e = r_1i_1 + r_2i_2 = r_3i_3 + r_4i_4$.

Les intensités sont donc solutions du système linéaire :

$$\begin{cases} r_1i_1 + r_2i_2 = e \\ r_3i_3 + r_4i_4 = e \\ r_1i_1 - r_4i_4 - r_5i_5 = 0 \\ r_2i_2 - r_3i_3 + r_5i_5 = 0 \\ i_1 - i_2 - i_3 + i_4 = 0 \\ i_3 - i_4 + i_5 = 0 \\ i_1 - i_2 + i_5 = 0 \end{cases}$$

Système homogène

Si le second membre b est le vecteur nul, on dit que le système linéaire (ou que l'équation linéaire correspondante) est *homogène*.

Soit $x_1u_1 + x_2u_2 + \cdots + x_nu_n = 0$ une équation linéaire homogène. Le vecteur nul $x_1 = x_2 = \cdots = x_n = 0$ est évidemment solution. Si $(x_1, x_2, \ldots, x_n)$ et $(x'_1, x'_2, \ldots, x'_n)$ sont solutions, alors on a

$$(x_1+x'_1)u_1+(x_2+x'_2)u_2+\cdots+(x_n+x'_n)u_n = \\ = (x_1u_1+x_2u_2+\cdots+x_nu_n)+(x'_1u_1+x'_2u_2+\cdots+x'_nu_n) = 0$$

et pour tout scalaire a,

$$(ax_1)u_1 + (ax_2)u_2 + \cdots + (ax_n)u_n = a(x_1u_1 + x_2u_2 + \cdots + x_nu_n) = 0\,.$$

La somme de deux solutions est donc une solution et le produit d'une solution par un scalaire est une solution.

Proposition. *L'ensemble des solutions d'un système linéaire homogène à n inconnues est un sous-espace vectoriel de $\mathbb{K}^n$.*

1.4 Vecteurs indépendants et bases

Définition

Soient $u_1, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$. Si l'équation linéaire $x_1u_1 + \cdots + x_nu_n = 0$ a pour seule solution $x_1 = x_2 = \cdots = x_n = 0$, on dit que les vecteurs $u_1, u_2, \ldots, u_n$ sont *indépendants*.

D'après cette définition, des vecteurs sont indépendants si aucun d'entre eux n'est combinaison linéaire des autres.

Théorème. *Soient $u_1,u_2,\ldots,u_n,v$ des vecteurs de $\mathbb{K}^p$. Si $u_1,u_2,\ldots,u_n$ sont indépendants et si $u_1,u_2,\ldots,u_n,v$ ne le sont pas, alors v est combinaison linéaire des vecteurs $u_1,u_2,\ldots,u_n$.*

Démonstration. Puisque les vecteurs $u_1,u_2,\ldots,u_n,v$ ne sont pas indépendants, il existe une relation linéaire $x_1u_1+x_2u_2+\cdots+x_nu_n+yv=0$ où les scalaires $x_1,\ldots,x_n,y$ ne sont pas tous nuls. Si l'on a $y=0$, alors il vient $x_1u_1+x_2u_2+\cdots+x_nu_n=0$, donc $x_1=x_2=\cdots=x_n=0$ car les vecteurs $u_1,u_2,\ldots,u_n$ sont indépendants ; dans ce cas, tous les scalaires $x_1,\ldots,x_n,y$

sont nuls et cela est contraire à ce qu'on a supposé. On en déduit que $y \neq 0$. De l'égalité de départ, on tire alors $v = -(x_1/y)u_1 - (x_2/y)u_2 - \cdots - (x_n/y)u_n$ et le vecteur v est bien combinaison linéaire de $u_1, \ldots, u_n$. ■

Théorème. *Soient $u_1, u_2, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$. Supposons qu'il existe des scalaires x_i et y_i tels que $x_1u_1 + x_2u_2 + \cdots + x_nu_n = y_1u_1 + y_2u_2 + \cdots + y_nu_n$. Si les vecteurs $u_1, u_2, \ldots, u_n$ sont indépendants, alors $x_i = y_i$ pour tout i.*

Démonstration. D'après les règles de calcul sur les vecteurs, on a

$$0=(x_1u_1+x_2u_2+\cdots+x_nu_n)-(y_1u_1+y_2u_2+\cdots+y_nu_n)=(x_1-y_1)u_1+(x_2-y_2)u_2+\cdots+(x_n-y_n)u_n$$

Si les vecteurs $u_1, u_2, \ldots, u_n$ sont indépendants, alors par définition $x_1 - y_1 = x_2 - y_2 = \cdots = x_n - y_n = 0$, donc $x_i = y_i$ pour tout i. ■

Définition
Soient $u_1, u_2, \ldots, u_n$ des vecteurs de $\mathbb{K}^p$ et soit V le sous-espace vectoriel engendré par ces vecteurs. Si $u_1, u_2 \ldots, u_n$ sont indépendants, on dit qu'ils forment une *base* de V.

D'après les théorèmes qu'on vient de montrer, on a ainsi la propriété suivante.

Des vecteurs $u_1, u_2, \ldots, u_n$ de $\mathbb{K}^p$ forment une base de V si et seulement si tout vecteur de V s'écrit de manière unique comme combinaison linéaire de $u_1, u_2, \ldots, u_n$.

Proposition. *Soient $u_1, u_2, \ldots, u_n$ des vecteurs appartenant à V. Ces vecteurs forment une base de V si et seulement si l'application $(x_1, x_2, \cdots, x_n) \mapsto x_1u_1 + x_2u_2 + \cdots + x_nu_n$ de $\mathbb{K}^n$ dans V est une bijection.*

Exemple. Les vecteurs canoniques $\boldsymbol{e_1}, \ldots, \boldsymbol{e_p}$ de $\mathbb{K}^p$ forment une base de $\mathbb{K}^p$.

En effet, tout vecteur $x=(x_1,\ldots,x_p)$ appartenant à $\mathbb{K}^p$ s'écrit $x_1\boldsymbol{e_1}+x_2\boldsymbol{e_2}+\cdots+x_p\boldsymbol{e_p}=x$. Tout vecteur de $\mathbb{K}^p$ s'écrit donc de manière unique comme combinaison des vecteurs canoniques.

Définition
La base de $\mathbb{K}^p$ formée des vecteurs canoniques $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_p}$ s'appelle la *base canonique* de $\mathbb{K}^p$.

Exemple. Résolvons l'équation linéaire $x_1u_1 + x_2u_2 = b$, où u_1, u_2 et b sont les vecteurs de $\mathbb{R}^2$ définis par $u_1 = (1, 1)$, $u_2 = (-2, -3)$ et $b = (b_1, b_2)$.
On a $x_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} -2 \\ -3 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_2 \\ x_1 - 3x_2 \end{bmatrix}$, donc l'équation s'écrit sous la forme du système linéaire

$$\begin{cases} x_1 - 2x_2 = b_1 \\ x_1 - 3x_2 = b_2 \end{cases}$$

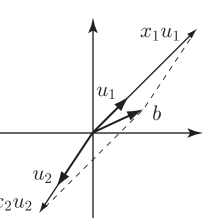
De la première équation, on tire $x_1 = b_1 + 2x_2$ et en reportant dans la deuxième équation, on obtient $(b_1 + 2x_2) - 3x_2 = -x_2 + b_1 = b_2$, d'où les équivalences

$$\begin{cases} x_1 - 2x_2 = b_1 \\ x_1 - 3x_2 = b_2 \end{cases} \iff \begin{cases} x_1 = b_1 + 2x_2 \\ x_2 = b_1 - b_2 \end{cases} \iff \begin{cases} x_1 = b_1 + 2(b_1 - b_2) \\ x_2 = b_1 - b_2 \end{cases}$$

Il y a une unique solution $(x_1, x_2) = (3b_1 - 2b_2, b_1 - b_2)$.
La seule façon d'exprimer le vecteur b comme combinaison linéaire des vecteurs u_1, u_2 consiste donc à écrire

$$(3b_1 - 2b_2)u_1 + (b_1 - b_2)u_2 = b\,.$$

Les vecteurs u_1,u_2 forment une base de $\mathbb{R}^2$. Si b=(0,0), l'équation est homogène et a pour unique solution $x_1 = x_2 = 0$.



2. Résolution des équations linéaires

2.1 Méthode de Gauss : pratique et exemples

Considérons le système d'équations

$$(S) \qquad \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 & (\mathrm{eq}_1) \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 & (\mathrm{eq}_2) \\ \qquad\vdots \qquad\qquad\qquad\quad \vdots & \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = b_p & (\mathrm{eq}_p) \end{cases}$$

Si le coefficient a_{11} n'est pas nul, on peut tirer x_1 de la première équation :

$$x_1 = b_1 - (a_{12}/a_{11})x_2 - (a_{13}/a_{11})x_3 - \cdots - (a_{1n}/a_{11})x_n\,.$$

En reportant cette expression de x_1 dans les autres équations, on obtient un système où ne figurent plus que les $n-1$ inconnues $x_2, \ldots, x_n$. La *méthode de Gauss* consiste à répéter cette opération successivement sur les inconnues restantes. C'est ainsi que nous avons procédé dans l'exemple précédent.

Pratiquement, on transforme le système (S) en un système équivalent (c'est-à-dire ayant les mêmes solutions) au moyen d'opérations sur les équations du système.

Opérations permises

opération 1 : changer l'ordre des équations ;

opération 2 : multiplier une équation eq_i par un scalaire a non nul, c'est-à-dire remplacer eq_i par $\mathrm{eq}'_i = a\ \mathrm{eq}_i$, où $a \neq 0$.

opération 3 : ajouter à une équation eq_i un multiple quelconque d'une autre, c'est-à-dire remplacer eq_i par $\mathrm{eq}'_i = \mathrm{eq}_i + a\,\mathrm{eq}_j$, où $i \neq j$ et où a est un scalaire quelconque.

Ces mêmes opérations permettent de retrouver les équations de départ : en effet,

- si $\mathrm{eq}'_i = a\ \mathrm{eq}_i$, où a est un scalaire non nul, alors $\mathrm{eq}_i = (1/a)\ \mathrm{eq}'_i$;
- si $\mathrm{eq}'_i = \mathrm{eq}_i + a\ \mathrm{eq}_j$, où $i \neq j$, alors $\mathrm{eq}_i = \mathrm{eq}'_i - a\ \mathrm{eq}_j$.

Lorsqu'on effectue ces opérations, le système d'équations est donc bien transformé en un système qui a les mêmes solutions.

Exemple 1. Soit le système d'équations $(S)\ \begin{cases} x_1 + 2x_2 - x_3 = b_1 & (\text{eq}_1) \\ 2x_1 + 5x_2 + x_3 = b_2 & (\text{eq}_2) \\ x_1 + 4x_2 + 5x_3 = b_3 & (\text{eq}_3) \\ x_1 + x_2 - 3x_3 = b_4 & (\text{eq}_4) \end{cases}$

Pour éliminer l'inconnue x_1 dans les équations eq_2, eq_3 et eq_4, formons les équations

$$\begin{cases} \text{eq}'_2 = \text{eq}_2 - 2\ \text{eq}_1 & : \quad x_2 + 3x_3 = b_2 - 2b_1 \\ \text{eq}'_3 = \text{eq}_3 - \text{eq}_1 & : \quad 2x_2 + 6x_3 = b_3 - b_1 \\ \text{eq}'_4 = \text{eq}_4 - \text{eq}_1 & : \quad -x_2 - 2x_3 = b_4 - b_1 \end{cases}$$

On obtient le système équivalent :

$$(S) \iff \begin{cases} x_1 + 2x_2 - x_3 = b_1 & (\text{eq}_1) \\ x_2 + 3x_3 = b_2 - 2b_1 & (\text{eq}'_2) \\ 2x_2 + 6x_3 = b_3 - b_1 & (\text{eq}'_3) \\ -x_2 - 2x_3 = b_4 - b_1 & (\text{eq}'_4) \end{cases}$$

Éliminons maintenant x_2 dans les deux dernières équations en formant les équations $\text{eq}''_3 = \text{eq}'_3 - 2\ \text{eq}'_2$ et $\text{eq}''_4 = \text{eq}'_4 + \text{eq}'_2$. Il vient

$$(S) \iff \begin{cases} x_1 + 2x_2 - x_3 = b_1 & (\text{eq}_1) \\ x_2 + 3x_3 = b_2 - 2b_1 & (\text{eq}'_2) \\ x_3 = -3b_1 + b_2 + b_4 & (\text{eq}''_4) \\ 0 = 3b_1 - 2b_2 + b_3 & (\text{eq}''_3) \end{cases}$$

Premier cas : $3b_1 - 2b_2 + b_3 \neq 0$. Le système (S) n'a pas de solution, car l'égalité (eq''_3) n'est pas satisfaite.

Second cas : $3b_1 - 2b_2 + b_3 = 0$. Le système se réduit aux trois équations (eq_1), (eq'_2), (eq''_4). La dernière donne x_3, puis on tire x_2 de la seconde et enfin x_1 de la première :

$$(S) \iff \begin{cases} x_1 = -2x_2 + x_3 + b_1 \\ x_2 = b_2 - 2b_1 - 3x_3 \\ x_3 = -3b_1 + b_2 + b_4 \end{cases} \iff \begin{cases} x_1 = -2x_2 + x_3 + b_1 \\ x_2 = 7b_1 - 2b_2 - 3b_4 \\ x_3 = -3b_1 + b_2 + b_4 \end{cases} \iff \begin{cases} x_1 = -16b_1 + 5b_2 + 7b_4 \\ x_2 = 7b_1 - 2b_2 - 3b_4 \\ x_3 = -3b_1 + b_2 + b_4 \end{cases}$$

et dans ce cas, le système a une unique solution.

Si la condition $3b_1 - 2b_2 - b_3 = 0$ est satisfaite, le système a une unique solution ; sinon, il n'a pas de solution.

Exemple 2. Modifions le système d'équations de l'exemple précédent en supprimant la dernière équation et en posant $b_1 = b_2 = 1$, $b_3 = -1$: l'égalité $3b_1 - 2b_2 + b_3 = 0$ est ainsi satisfaite. Ce système s'écrit

$$(S)\ \begin{cases} x_1 + 2x_2 - x_3 = 1 & (\text{eq}_1) \\ 2x_1 + 5x_2 + x_3 = 1 & (\text{eq}_2) \\ x_1 + 4x_2 + 5x_3 = -1 & (\text{eq}_3) \end{cases}$$

En faisant les mêmes opérations que précédemment, il vient les équivalences

$$(S) \iff \begin{cases} x_1+2x_2-x_3 = 1 & (\mathrm{eq}_1) \\ \quad\; x_2+3x_3 = -1 & (\mathrm{eq}_2') \end{cases} \iff \begin{cases} x_1 = 1-2x_2+x_3 \\ x_2 = -1-3x_3 \end{cases} \iff \begin{cases} x_1 = 3+7x_3 \\ x_2 = -1-3x_3 \end{cases}$$

Si l'on donne à x_3 n'importe quelle valeur t, on obtient les solutions

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3+7t \\ -1-3t \\ t \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 7 \\ -3 \\ 1 \end{bmatrix}$$

A des valeurs de t différentes correspondent des solutions différentes : le système (S) possède donc une infinité de solutions. Si l'on représente (x_1, x_2, x_3) par un point de l'espace $\mathbb{R}^3$, les solutions décrivent la droite passant par le point $(3, -1, 0)$ et de vecteur directeur $(7, -3, 1)$.

Exemple 3. Soit le système linéaire homogène $(S) \begin{cases} 2x_1 - x_2 + x_3 - x_4 = 0 & (\mathrm{eq}_1) \\ 3x_1 - x_2 + 3x_3 - 2x_4 = 0 & (\mathrm{eq}_2) \end{cases}$ où les inconnues sont les nombres réels x_1, x_2, x_3, x_4.

En remplaçant eq_2 par $2\,\mathrm{eq}_2 - 3\,\mathrm{eq}_1$, l'inconnue x_1 disparaît dans la deuxième équation et (S) est équivalent à

$$\begin{cases} 2x_1-x_2+x_3-x_4 = 0 \ (\mathrm{eq}_1) \\ \quad\; x_2+3x_3-x_4 = 0 \ (2\ \mathrm{eq}_2 - 3\ \mathrm{eq}_1)) \end{cases} \iff \begin{cases} 2x_1 = x_2-x_3+x_4 \\ \ x_2 = -3x_3+x_4 \end{cases} \iff \begin{cases} x_1 = -2x_3+x_4 \\ x_2 = -3x_3+x_4 \end{cases}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2x_3 + x_4 \\ -3x_3 + x_4 \\ x_3 \\ x_4 \end{bmatrix} = x_3 \begin{bmatrix} -2 \\ -3 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

On peut donner à x_3 et x_4 des valeurs réelles arbitraires : les solutions de (S) sont donc

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = t \begin{bmatrix} -2 \\ -3 \\ 1 \\ 0 \end{bmatrix} + t' \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \quad \text{où } t \text{ et } t' \text{ sont des nombres quelconques.}$$

Posons $u = (-2, -3, 1, 0)$ et $u' = (1, 1, 0, 1)$. Les solutions sont toutes les combinaisons linéaires de u et u', autrement dit l'ensemble des solutions est le sous-espace vectoriel V de $\mathbb{R}^4$ engendré par les vecteurs u, u'. Les deux dernières coordonnées du vecteur $tu + t'u'$ sont t et t' ; si l'on suppose que $tu + t'u' = 0$, alors $t = 0$ et $t' = 0$. Cela montre que les vecteurs u et u' sont indépendants. Les vecteurs u et u' forment donc une base du sous-espace vectoriel V. Le système possède une infinité de solutions.

Remarque

La méthode de Gauss n'est pas une bonne méthode numérique pour résoudre des systèmes linéaires de grande taille : elle nécessite beaucoup d'opérations et le résultat est trop sensible aux erreurs d'arrondi. Nous présenterons au chapitre 8 quelques méthodes de résolution numérique plus efficaces.

2.2 Résolution des systèmes linéaires

Reprenons le système de p équations à n inconnues

$$(S) \qquad \begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 & (\mathrm{eq}_1) \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 & (\mathrm{eq}_2) \\ \qquad\vdots \qquad\qquad \vdots & \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = b_p & (\mathrm{eq}_p) \end{cases}$$

Échelonnement du système. Supposons que dans (S), l'un des coefficients de x_1 n'est pas nul (sinon, l'inconnue x_1 n'apparaît pas et le système n'a que les $n-1$ inconnues $x_2, \ldots, x_n$). En plaçant en première position l'une des équations où apparaît x_1, on peut supposer $a_{11} \neq 0$.
Comme on l'a vu, la méthode de Gauss consiste à faire les opérations qui éliminent x_1 dans les équations $\mathrm{eq}_2, \ldots, \mathrm{eq}_p$, et à continuer successivement l'élimination des inconnues qui restent.
À chaque étape, on a éliminé au moins une inconnue dans les dernières équations ; cela peut produire certaines équations où ne figure plus d'inconnue, comme dans l'exemple 1.
En réordonnant éventuellement les équations en cours de calcul, on trouve, comme dans les exemples précédents, un système (S') équivalent à (S) et de la forme :

$$(S') \qquad \begin{cases} x_1 + \quad a'_{12}x_2 + a'_{13}x_3 + \cdots + a'_{1n}x_n = b'_1 & (\mathrm{eq}'_1) \\ \quad x_{i_2} + \quad a'_{2\,i_2+1}x_{i_2+1} + \cdots + a'_{2n}x_n = b'_2 & (\mathrm{eq}'_2) \\ \qquad x_{i_3} + \quad a'_{3\,i_3+1}x_{i_3+1} + \cdots + a'_{3n}x_n = b'_3 & (\mathrm{eq}'_3) \\ \qquad\quad \ddots \qquad\qquad \vdots \qquad\qquad \vdots & \\ \qquad\qquad x_{i_r} + \; a'_{r\,i_r+1}x_{i_r+1} + \cdots + a'_{r\,n}x_n = b'_r & (\mathrm{eq}'_r) \\ \qquad\qquad\qquad\qquad\qquad 0 = b'_{r+1} & (\mathrm{eq}'_{r+1}) \\ \qquad\qquad\qquad\qquad\qquad \vdots \quad \vdots & \\ \qquad\qquad\qquad\qquad\qquad 0 = b'_p & (\mathrm{eq}'_p) \end{cases}$$

Un tel système est dit *en échelons*. Voici les caractéristiques d'un système en échelons.

i) Dans les r premières équations, les *inconnues de tête* $x_1, x_{i_2}, x_{i_3}, \ldots, x_{i_r}$ ont leurs indices strictement croissants : $1 = i_1 < i_2 < i_3 < \cdots < i_r \leqslant n$.

ii) Chaque inconnue de tête a pour coefficient 1 : on obtient cela en divisant une équation $\alpha x_k + \alpha_{k+1}x_{k+1} + \cdots + \alpha_n x_n = \beta$ par le coefficient $\alpha \neq 0$ de l'inconnue de tête.

iii) Le nombre r d'équations contenant des inconnues est bien sûr inférieur ou égal à n et à p.

Nous verrons bientôt que ce nombre r d'inconnues de tête ne dépend que du système (S) et non pas de la manière dont on s'y est pris pour l'échelonner.

Définitions

- Le nombre r s'appelle le *rang* du système (S) ou de l'équation linéaire.
- Si l'on a $r < p$, les $p - r$ dernières égalités (celles dont le premier membre est nul) s'appellent les *égalités de compatibilité*.

Puisque les systèmes (S) et (S') ont les mêmes solutions, on en déduit :

Si l'une des égalités de compatibilité n'est pas satisfaite, alors le système n'a pas de solution.

Compte-tenu du déroulement de la méthode de Gauss, on a les propriétés suivantes.

- Le rang d'un système linéaire ne dépend pas du second membre.
- Le second membre de (S') ne dépend que du second membre de (S).
- Si le système est homogène (son second membre est nul), alors toutes les égalités de compatibilité sont satisfaites : en effet, chaque nombre b'_i s'obtient en faisant des combinaisons linéaires de $b_1, \ldots, b_r$; si tous les b_i sont nuls, il en va donc de même des b'_i.

Fin de la résolution. Supposons que l'on a $r = p$ (il n'y a pas d'égalité de compatibilité) ou bien que toutes les égalités de compatibilité sont satisfaites : dans ces cas, le système se réduit aux r premières équations.

On poursuit alors la résolution en exprimant, dans les r premières équations, les inconnues de tête en fonction des autres ; pour simplifier, supposons que les inconnues de tête sont $x_1, x_2, \ldots, x_r$ (il suffit de rénuméroter les inconnues pour qu'il en soit ainsi). Le système (S') s'écrit alors sous la forme

$$(S^*) \qquad \begin{cases} x_1 = b_1^* - a_{12}^* x_2 - \quad \cdots \quad \cdots \quad - a_{1n}^* x_n \\ x_2 = b_2^* \qquad - a_{23}^* x_3 - \quad \cdots \quad - a_{2n}^* x_n \\ \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ x_r = b_r^* \qquad\qquad - a_{r\,r+1}^* x_{r+1} - \cdots - a_{rn}^* x_n \end{cases}$$

Il est maintenant facile d'exprimer chacune des inconnues $x_1, \ldots, x_r$ au moyen de $x_{r+1}, \ldots, x_n$:

- on reporte l'expression de x_r dans l'avant dernière équation, ce qui donne x_{r-1} en fonction de $x_{r+1}, \ldots, x_n$,
- puis on opère par reports successifs dans les expressions de $x_{r-2}, \ldots, x_1$.

Distinguons deux cas.

Premier cas : $r = n$. Alors la dernière équation s'écrit $x_n = b_n^*$ et de proche en proche, on calcule la valeur de $x_{n-1}, \ldots, x_1$. Le système possède dans ce cas une unique solution.

Second cas : $r < n$. Après report des inconnues $x_{r+1}, \ldots, x_n$, le système s'écrit :

$$\begin{cases} x_1 = \beta_1 + \gamma_{1\,r+1} x_{r+1} + \gamma_{1\,r+2} x_{r+2} + \ldots + \gamma_{1n} x_n \\ x_2 = \beta_2 + \gamma_{2\,r+1} x_{r+1} + \gamma_{2\,r+2} x_{r+2} + \ldots + \gamma_{2n} x_n \\ \vdots \qquad\qquad \vdots \\ x_r = \beta_r + \gamma_{r\,r+1} x_{r+1} + \gamma_{r\,r+2} x_{r+2} + \ldots + \gamma_{rn} x_n \end{cases}$$

ou encore

$$\begin{bmatrix} x_1 \\ \vdots \\ x_r \\ x_{r+1} \\ x_{r+2} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_{r+1} \begin{bmatrix} \gamma_{1\,r+1} \\ \vdots \\ \gamma_{r\,r+1} \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_{r+2} \begin{bmatrix} \gamma_{1\,r+2} \\ \vdots \\ \gamma_{r\,r+2} \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_n \begin{bmatrix} \gamma_{1n} \\ \vdots \\ \gamma_{rn} \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

Posons $x=(x_1,\ldots,x_n)$, $w=(\beta_1,\ldots,\beta_r,0,\ldots,0)$ et notons $w_{r+1},\ldots,w_n$ les vecteurs en facteur des inconnues $x_{r+1},\ldots,x_n$. On a donc

$$x = w + x_{r+1}w_{r+1} + \cdots + x_n w_n\,.$$

- On obtient les solutions du système en donnant des valeurs arbitraires aux inconnues $x_{r+1},\ldots,x_n$. Les solutions x s'obtiennent donc en ajoutant le vecteur w à n'importe quelle combinaison linéaire des $n-r$ vecteurs $w_{r+1},\ldots,w_n$. Le système possède une infinité de solutions.
- Les vecteurs $w_{r+1},\ldots,w_n$ sont indépendants : en effet, si $t_{r+1},\ldots,t_n$ sont des scalaires, les $n-r$ dernières coordonnées du vecteur $t_{r+1}w_{r+1}+\cdots+t_nw_n$ sont $t_{r+1},\ldots,t_n$; si l'on a $t_{r+1}w_{r+1}+\cdots+t_nw_n=0$, il vient donc $t_{r+1}=\cdots=t_n=0$.

Conclusion : Les solutions s'écrivent de manière unique

$$x = w + t_{r+1}w_{r+1} + \cdots + t_nw_n, \quad \text{où } t_{r+1},\ldots,t_n \text{ sont des scalaires quelconques.}$$

Les nombres $t_{r+1},\ldots,t_n$ s'appellent les *paramètres* des solutions.

Cas d'un système homogène. Si le système (S) est homogène ($b=0$), les nombres b_i^* sont tous nuls, donc $w=0$. L'ensemble des solutions est le sous-espace vectoriel V de $\mathbb{K}^p$ engendré par les vecteurs $w_{r+1},\ldots,w_n$; puisque ces vecteurs sont indépendants, ils forment une base de V.

Équations d'un sous-espace vectoriel

Soit V le sous-espace vectoriel de $\mathbb{K}^p$ engendré par des vecteurs $u_1,u_2,\ldots,u_n$. Un vecteur y appartient à V si et seulement si l'équation linéaire

$$(*) \qquad x_1u_1 + x_2u_2 + \cdots + x_nu_n = y$$

a au moins une solution, ce qui a lieu si et seulement si les égalités de compatibilités sont satisfaites : ces égalités constituent donc des équations du sous-espace vectoriel V.

Exemple. Soit V le sous-espace vectoriel de $\mathbb{R}^4$ engendré par les vecteurs $u_1=(1,1,2,1)$, $u_2=(-1,1,1,0)$, $u_3=(1,1,2,1)$. Un vecteur $(y_1,y_2,y_3,y_4)\in\mathbb{R}^4$ est combinaison linéaire de u_1,u_2,u_3 si et seulement si le système linéaire suivant a au moins

une solution :
$$\begin{cases} x_1 - x_2 + x_3 = y_1 & (\mathrm{eq}_1) \\ x_1 + x_2 + x_3 = y_2 & (\mathrm{eq}_2) \\ 2x_1 + x_2 + 2x_3 = y_3 & (\mathrm{eq}_3) \\ x_1 + x_3 = y_4 & (\mathrm{eq}_4) \end{cases}$$
On a les systèmes équivalents :
$$\begin{cases} x_1 - x_2 + x_3 = y_1 & (\mathrm{eq}_1) \\ 2x_2 = -y_1 + y_2 & (\mathrm{eq}_2')=(\mathrm{eq}_2)-(\mathrm{eq}_1) \\ 3x_2 = -2y_1 + y_3 & (\mathrm{eq}_3')=(\mathrm{eq}_3)-2(\mathrm{eq}_1) \\ x_2 = -y_1 + y_4 & (\mathrm{eq}_4')=(\mathrm{eq}_4)-(\mathrm{eq}_1) \end{cases}$$
$$\iff \begin{cases} x_1 - x_2 + x_3 = y_1 & (\mathrm{eq}_1) \\ 2x_2 = -y_1 + y_2 & (\mathrm{eq}_2') \\ 0 = -2(-2y_1 + y_3) + 3(-y_1 + y_2) & -2(\mathrm{eq}_3')+3(\mathrm{eq}_2') \\ 0 = -2(-y_1 + y_4) + (-y_1 + y_2) & -2(\mathrm{eq}_4')+(\mathrm{eq}_2') \end{cases}$$
Pour qu'il y ait une solution (x_1, x_2, x_3), il faut et il suffit que les égalités de compatibilité soient satisfaites. Le sous-espace vectoriel V a donc pour équations $\begin{cases} y_1 + 3y_2 - 2y_3 = 0 \\ y_1 + y_2 - 2y_4 = 0 \end{cases}$.

- Par exemple, le vecteur $(5, 1, 4, 3)$ appartient à V et $(0, 4, 6, 1)$ n'y appartient pas.
- Soient $v_1 = (0, 1, 1, 0)$ et $v_2 = (1, 1, -1, 1)$. Quelles sont les combinaisons de v_1 et v_2 qui appartiennent à V ? Les coordonnées d'un vecteur y combinaison de v_1 et v_2 sont de la forme $\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = t \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + t' \begin{bmatrix} 1 \\ 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} t' \\ t + t' \\ t - t' \\ t' \end{bmatrix}$. On a $y_1 + 3y_2 - 2y_3 = t' + 3(t + t') - 2(t - t') = t + 6t'$ et $y_1 + y_2 - 2y_4 = t' + (t + t') - 2t' = t$. Le vecteur y appartient à V si et seulement si l'on a $t + 6t' = t = 0$, c'est-à-dire si et seulement si $t = t' = 0$. Par conséquent, une combinaison linéaire de v_1 et v_2 n'appartient à V que si c'est le vecteur nul.

2.3 Les résultats généraux

La méthode de Gauss donne les principaux renseignement sur les solutions d'un système linéaire.

Résultats 1. *Supposons que le système linéaire (S) possède p équations, n inconnues et que son rang est $r \geqslant 1$.*

a) *Si $r < n$ et s'il y a au moins une solution, alors il y a une infinité de solutions.*

b) *Si $p < n$ et s'il y a au moins une solution, alors il y a une infinité de solutions.*

c) *Si l'on a $r = p < n$, alors le système possède une infinité de solutions.*

Concernant les systèmes homogènes, on a les résultats suivants.

Résultats 2. *Supposons que le système linéaire* (S) *est homogène.*

a) *Si* $r < n$, *l'ensemble des solutions du système est un sous-espace vectoriel de* $\mathbb{K}^n$ *ayant une base formée de* $n-r$ *vecteurs.*

b) *Le système a une unique solution si et seulement si* $r = n$ *et dans ce cas, la solution est le vecteur nul.*

c) *Si* $p < n$, *le système a une infinité de solutions.*

La proposition suivante concerne le cas important des systèmes linéaires ayant autant d'équations que d'inconnues ($p = n$). Sous forme vectorielle, un tel système s'écrit

$$x_1 u_1 + \cdots + x_n u_n = b, \quad \text{où } u_1, \ldots, u_n \text{ et } b \text{ sont des vecteurs de } \mathbb{K}^n.$$

Résultat 3. *Soient* $u_1, \ldots, u_n$ *des vecteurs de* $\mathbb{K}^n$*. Les conditions suivantes sont équivalentes :*

a) *le rang du système linéaire* $x_1 u_1 + \cdots + x_n u_n = 0$ *est égal à* n *;*

b) *les vecteurs* $u_1, \ldots, u_n$ *sont indépendants ;*

c) *quel que soit le vecteur* $b \in \mathbb{K}^n$, *l'équation linéaire* $x_1 u_1 + \cdots + x_n u_n = b$ *a une unique solution ;*

d) *l'application* $(x_1, \ldots, x_n) \mapsto x_1 u_1 + \cdots + x_n u_n$ *de* $\mathbb{K}^n$ *dans* $\mathbb{K}^n$ *est bijective.*

Démonstrations des résultats. Nous utilisons la discussion et les notations du paragraphe précédent, pages 111-112.

1. S'il y a au moins une solution, c'est que les éventuelles égalités de compatibilité sont satisfaites. Si de plus $r < n$, alors on est dans le second cas et il y a une infinité de solutions ; si $p < n$, alors on a $r < n$, car $r \leqslant p$: cela montre (b). Si $r = p < n$, il n'y a pas d'égalité de compatibilité et l'on est encore dans le second cas.

2. Puisque le système est homogène, les égalités de compatibilité sont toutes satisfaites. Si $r < n$, on est dans le second cas et nous savons que le vecteur w est nul. Les solutions sont donc les combinaisons linéaires des $n-r$ vecteurs indépendants $w_{r+1}, \ldots, w_n$. Puisqu'on a toujours $r \leqslant n$, on en déduit que si le système n'a qu'une solution, alors $r = n$. Réciproquement, si $r = n$, alors on est dans le premier cas et il y a donc une unique solution qui est nécessairement le vecteur nul. Nous avons ainsi montré les propriétés (a) et (b). La propriété (c) résulte de (a) : en effet, on a toujours $r \leqslant p$, donc si $p < n$, alors on a $r < n$.

3. Les vecteurs $u_1, \ldots, u_n$ sont indépendants si et seulement si l'équation $x_1 u_1 + \cdots + x_n u_n = 0$ a pour seule solution $x_1 = \cdots = x_n = 0$. D'après le résultat 2 (b), cela équivaut à dire que le rang de cette équation linéaire est n, d'où l'équivalence entre (a) et (b).
Soit $b \in \mathbb{K}^n$. Les systèmes linéaires $x_1 u_1 + \cdots + x_n u_n = b$ et $x_1 u_1 + \cdots + x_n u_n = 0$ on même rang. Si les vecteurs $u_1, \ldots, u_n$ sont indépendants, le système linéaire $x_1 u_1 + \cdots + x_n u_n = b$ a donc pour rang n. On est alors dans le premier cas et le système a une unique solution. Réciproquement, si le système linéaire $x_1 u_1 + \cdots + x_n u_n = b$ a une unique solution, alors d'après le résultat 1 (a), l'inégalité $r < n$ n'est pas vraie ; puisqu'on a $r \leqslant n$, il s'ensuit $r = n$. Nous avons ainsi montré l'équivalence entre (a) et (c).
Les propriétés (c) et (d) sont équivalentes par définition d'une application bijective. ■

3. Dimension d'un sous-espace vectoriel

3.1 Définition de la dimension

Voici une conséquence très importante du résultat 3 précédent.

Proposition. *Soient $u_1, \ldots, u_m$ des vecteurs indépendants appartenant à $\mathbb{K}^p$. Supposons que $v_1, \ldots, v_q$ sont des vecteurs de $\mathbb{K}^p$ et que chaque v_i est combinaison linéaire de $u_1, \ldots, u_m$. Si $q > m$, les vecteurs $v_1, \ldots, v_q$ ne sont pas indépendants.*

Démonstration. Par hypothèse, chaque vecteur v_i s'écrit $v_i = a_{1i}u_1 + a_{2i}u_2 + \cdots + a_{mi}u_m$, où les a_{ji} sont des scalaires. Soient $x_1, \ldots, x_q$ des scalaires et soit $e = x_1v_1 + \cdots + x_qv_q$. On a $x_iv_i = x_ia_{1i}u_1 + x_ia_{2i}u_2 + \cdots + x_ia_{mi}u_m$, donc le coefficient de u_1 dans la combinaison linéaire e est $a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$. De même, le coefficient de u_i dans e est $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{iq}x_q$. Supposons $e = 0$. Puisque $u_1, \ldots, u_m$ sont indépendants, il vient

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q & = 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q & = 0 \\ \qquad\qquad \vdots & \ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mq}x_q & = 0 \end{cases}$$

Si $q > m$, ce système a plus d'inconnues que d'équations et comme il est homogène, il a une infinité de solutions (résultat 2 (c)). L'équation linéaire $x_1v_1 + \cdots + x_qv_q = 0$ n'a pas que la solution nulle, donc les vecteurs $v_1, \ldots, v_q$ ne sont pas indépendants. ■

Corollaire. *Si $v_1, \ldots, v_q$ sont des vecteurs indépendants de $\mathbb{K}^p$, alors on a $q \leqslant p$.*

Démonstration. Tout vecteur de $\mathbb{K}^p$ est combinaison linéaire des vecteurs canoniques $\boldsymbol{e_1}, \ldots, \boldsymbol{e_p}$ qui sont indépendants. Chaque vecteur v_i est donc combinaison de $\boldsymbol{e_1}, \ldots, \boldsymbol{e_p}$. Si l'on a $q > p$, alors d'après la proposition précédente, les vecteurs $v_1, \ldots, v_q$ ne sont pas indépendants. Si les vecteurs $v_1, \ldots, v_q$ sont indépendants, c'est donc que l'on a $q \leqslant p$. ■

Théorème. *Tout sous-espace vectoriel de $\mathbb{K}^p$ possède des bases.*

Démonstration. Soit V un sous-espace vectoriel de $\mathbb{K}^p$. D'après le corollaire précédent, toute partie de $\mathbb{K}^p$ formée de vecteurs indépendants possède au plus p éléments. Choisissons une partie $\{u_1,u_2,\ldots,u_d\}$ de V formée de vecteurs indépendants et ayant le plus grand nombre possible d'éléments. Soit $v \in V$. Si v est l'un des vecteurs u_i, alors v est combinaison linéaire de $u_1,\ldots,u_d$. Supposons que v est différent de tous les u_i. Alors l'ensemble $\{u_1, u_2, \ldots, u_d, v\}$ possède $d+1$ éléments, donc les vecteurs $u_1,u_2,\ldots,u_d,v$ ne sont pas indépendants. D'après le théorème page 105, v est combinaison linéaire de $u_1,u_2,\ldots,u_d$. Ainsi tout vecteur de V est combinaison linéaire de $u_1,\ldots,u_d$. Puisque ces vecteurs sont indépendants, ils forment par définition une base de V. ■

Proposition. *Soit V un sous-espace vectoriel de $\mathbb{K}^p$. Toutes les bases de V ont le même nombre d'éléments.*

Démonstration. Supposons que les vecteurs $u_1, \ldots, u_m$ forment une base de V et que les vecteurs $v_1,\ldots,v_q$ forment aussi une base de V. Chaque vecteur v_i appartient à V, donc est combinaison linéaire des vecteurs indépendants $u_1,\ldots,u_m$. Si l'on avait $q>m$, alors d'après la propo-

sition précédente, les vecteurs $v_1, \ldots, v_q$ ne seraient pas indépendants. On a donc $q \leqslant m$. En échangeant le rôle des bases dans ce raisonnement, on montre de même que l'on a $m \leqslant q$, d'où $m = q$. ■

Définition
Si V est un sous-espace vectoriel de $\mathbb{K}^p$, le nombre d'éléments d'une base de V s'appelle la *dimension* de V et se note $\dim V$.

Voici des conséquences immédiates de la définition.

a) Puisque la base canonique de $\mathbb{K}^p$ possède p éléments, on a $\dim \mathbb{K}^p = p$.

b) Si V est un sous-espace vectoriel de $\mathbb{K}^p$, alors $\dim V \leqslant p$.

En effet, une base de V étant formée de vecteurs indépendants, le nombre de ces vecteurs n'excède pas p.

c) Si un système linéaire homogène à n inconnues est de rang $r < n$, l'ensemble des solutions est un sous-espace vectoriel de dimension $n-r$ (résultats 2, page 114). Deux systèmes linéaires homogènes qui ont les mêmes solutions ont donc le même rang.

d) Soit V un sous-espace vectoriel de $\mathbb{K}^p$. Si $\dim V = d < p$, alors tout système d'équations définissant V est de rang $p-d$. Un système minimal d'équations définissant V possède donc $p-d$ équations.

3.2 Droites, plans, hyperplans

- Un sous-espace vectoriel de dimension 1 s'appelle une *droite* ; un sous-espace vectoriel de dimension 2 s'appelle un *plan*.
- Un sous-espace vectoriel de $\mathbb{K}^p$ de dimension $p-1$ s'appelle un *hyperplan de* $\mathbb{K}^p$.

Un hyperplan de $\mathbb{K}^p$ est donc défini par une seule équation (car $p - (p-1) = 1$).

Exemples

- Si a, b, c, d sont des nombres réels non tous nuls, le sous-espace vectoriel de $\mathbb{R}^4$ d'équation $ax + by + cz + dt = 0$ est un hyperplan de $\mathbb{R}^4$.
- Si a, b, c sont des scalaires non tous nuls, le sous-espace vectoriel de $\mathbb{R}^3$ d'équation $ax + by + cz = 0$ est un plan.
  Par exemple, le plan d'équation $x - 2z = 0$ a pour base les vecteurs $(2,0,1), (0,1,0)$.
- Pour une droite de $\mathbb{R}^3$, le nombre minimal d'équations est $3-1 = 2$: une droite de $\mathbb{R}^3$ est définie par deux équations non proportionnelles.
- Le sous-espace vectoriel de $\mathbb{R}^3$ d'équations $\begin{cases} x - y + z = 0 \\ 2x + y - z = 0 \end{cases}$ est la droite de vecteur directeur $(0, 1, 1)$; c'est l'intersection du plan d'équation $x - y + z = 0$ et du plan d'équation $2x + y - z = 0$.

Sous-espaces affines de $\mathbb{R}^3$

Définition

Soit V un sous-espace vectoriel de $\mathbb{R}^3$. Si A est un point de $\mathbb{R}^3$, le *sous-espace affine passant par* A *et de direction* V est l'ensemble des points $M \in \mathbb{R}^3$ tels que $\overrightarrow{AM} \in V$.

- Supposons par exemple que V est le plan d'équation $ax + by + cz = 0$ et que les coordonnées de A sont (x_0, y_0, z_0). Le plan affine passant par A et de direction V a pour équation $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$.
- Soit $u = (p, q, r)$ un vecteur non nul. La droite affine passant par (x_0, y_0, z_0) et de vecteur directeur u est l'ensemble des points $(x_0, y_0, z_0) + t(p, q, r)$, où t parcourt $\mathbb{R}$.

Plans en feuillets. Soient P et P' deux plans affines de $\mathbb{R}^3$, sécants selon une droite D. Les équations de ces plans sont (P) : $ax + by + cz + d = 0$, (P') : $a'x + b'y + c'z + d' = 0$.
Cherchons la condition pour qu'un plan Π d'équation $ux + vy + wz + h = 0$ contienne la droite D.
La droite D est l'ensemble des solutions du système linéaire

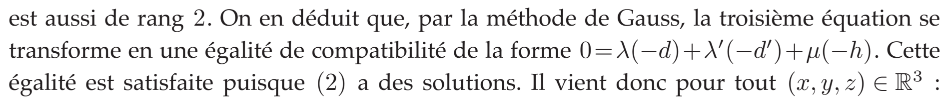
(1) $$\begin{cases} ax + by + cz &= -d \\ a'x + b'y + c'z &= -d' \end{cases}$$

qui est par hypothèse de rang 2.

L'intersection $\Pi \cap D$ a donc pour équations

(2) $$\begin{cases} ax + by + cz &= -d \\ a'x + b'y + c'z &= -d' \\ ux + vy + wz &= -h \end{cases}$$

Supposons que le plan Π contient la droite D. Alors les systèmes (1) et (2) ont les mêmes solutions, donc le système (2) est aussi de rang 2. On en déduit que, par la méthode de Gauss, la troisième équation se transforme en une égalité de compatibilité de la forme $0 = \lambda(-d) + \lambda'(-d') + \mu(-h)$. Cette égalité est satisfaite puisque (2) a des solutions. Il vient donc pour tout $(x, y, z) \in \mathbb{R}^3$:



$$\lambda(ax + by + cz + d) + \lambda'(a'x + b'y + c'z + d') + \mu(ux + vy + wz + h) = 0$$

On a $\mu \neq 0$, car les équations de P et P' ne sont pas proportionnelles. Par conséquent,

$$ux + vy + wz + h = -(\lambda/\mu)(ax + by + cz + d) - (\lambda'/\mu)(a'x + b'y + c'z + d'),$$

autrement dit : l'équation de Π est combinaison des équations de P et de P'.
Réciproquement, supposons que l'équation de Π est combinaison des équations de P et de P'. Tout point de D satisfait à la fois l'équation de P et celle de P', donc satisfait l'équation de Π : tout point de D est donc aussi dans Π.

Résumons : si P et P' sont des plans de $\mathbb{R}^3$ sécants selon une droite, les plans contenant cette droite sont ceux dont l'équation est combinaison des équations de P et de P'.

3.3 Recherche d'une base

Vecteurs en échelons

Définition

Soit $u_1, u_2, \ldots, u_n$ une suite de vecteurs de $\mathbb{K}^p$. On dit que la suite est *échelonnée* si ces vecteurs sont tous non nuls et si, pour tout $i = 2, 3, \ldots, n$, la première coordonnée non nulle de u_i est d'indice strictement supérieur à l'indice de la première coordonnée non nulle de u_{i-1}.

Exemple. Voici une suite échelonnée de vecteurs de $\mathbb{R}^5$:

$$u_1 = \begin{bmatrix} 2 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0 \\ 7 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}, \quad u_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ c_5 \end{bmatrix}, \quad u_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \end{bmatrix}$$

L'équation linéaire $x_1u_1 + x_2u_2 + x_3u_3 + x_4u_4 = 0$ s'écrit

$$\begin{cases} 2x_1 = 0 \\ a_2x_1 + 7x_2 = 0 \\ a_3x_1 + b_3x_2 = 0 \\ a_4x_1 + b_4x_2 + x_3 = 0 \\ a_5x_1 + b_5x_2 + c_5x_3 + 3x_4 = 0 \end{cases}$$

La première équation donne $x_1 = 0$, puis la deuxième s'écrit $a_2 \times 0 + 7x_2 = 0$, donc $x_2 = 0$; la troisième équation est satisfaite, de la quatrième on tire $x_3 = 0$, et enfin la dernière équation s'écrit $3x_4 = 0$, donc $x_4 = 0$. L'équation linéaire a pour seule solution $x_1 = x_2 = x_3 = x_4 = 0$, donc les vecteurs u_1, u_2, u_3, u_4 sont indépendants.

Ce raisonnement pouvant s'appliquer à une suite échelonnée quelconque, on a la proposition suivante.

Proposition. *Des vecteurs échelonnés sont indépendants.*

Recherche d'une base

Dans ce paragraphe, V est le sous-espace vectoriel de $\mathbb{K}^p$ engendré par des vecteurs $u_1, u_2, \ldots, u_n$.
On peut effectuer sur les vecteurs $u_1, \ldots, u_n$ les mêmes opérations que celles que l'on pratique sur les équations au cours de la méthode de Gauss.

opération (a) : changer l'ordre des vecteurs ;

opération (b) : multiplier un vecteur par un scalaire $a \neq 0$;

opération (c) : ajouter à l'un des vecteurs un multiple d'un autre, c'est-à-dire remplacer u_i par $u'_i = u_i + au_j$, où $i \neq j$ et où a est un scalaire quelconque.

Puisqu'on a $u_i = u'_i - au_j$, l'opération (c) permet de retrouver u_i à partir de u'_i.

Proposition. *Si l'on applique les opérations précédentes aux vecteurs $u_1, \ldots, u_n$, on obtient des vecteurs qui engendrent le même sous-espace vectoriel V.*

Démonstration. Posons $u'_1 = u_1 + bu_2$, où b est un scalaire quelconque. Le vecteur u'_1 est combinaison linéaire de u_1 et u_2, donc u'_1 appartient à V. Il s'ensuit que toute combinaison linéaire des vecteurs $u'_1, u_2, \ldots, u_n$ est dans V. De même, u_1 est combinaison linéaire de u'_1 et u_2 donc tout vecteur de V est combinaison linéaire de $u'_1, u_2, \ldots u_n$. ■

Nous allons voir que par les opérations (a), (b) et (c), on peut transformer les vecteurs $u_1, \ldots, u_n$ en une suite échelonnée.

Considérons un vecteur $u = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$ et supposons que sa k-ième coordonnée a_k est non nulle. Si $v = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}$ est un vecteur quelconque, le vecteur $v' = v - (b_k/a_k)u$ a pour k-ième coordonnée $b_k - (b_k/a_k)a_k = 0$ et s'obtient à partir de u, v par l'opération (c). Supposons par exemple que la première coordonnée de u_1 n'est pas nulle. En ajoutant à chacun des vecteurs $u_2, \ldots, u_n$ un multiple convenable de u_1, on peut ainsi obtenir des vecteurs $u'_2, \ldots, u'_n$ dont la première coordonnée est nulle. En poursuivant ces opérations, on formera des vecteurs en échelon.

Exemple. Échelonnons la suite de vecteurs

$$u_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 3 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 3 \\ -1 \\ 1 \\ 4 \end{bmatrix}, \quad u_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}, \quad u_4 = \begin{bmatrix} -2 \\ 4 \\ 1 \\ 1 \end{bmatrix}$$

Étape 1. En ajoutant $3u_1$ à $-2u_2$ (opérations (b) et (c)), on obtient le vecteur $u'_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}$ dont la première coordonnée est nulle ; de même, en posant $u'_3 = 2u_3 - u_1$ et $u'_4 = u_4 + u_1$, il vient

$$u_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 3 \end{bmatrix}, \quad u'_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}, \quad u'_3 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}, \quad u'_4 = \begin{bmatrix} 0 \\ 4 \\ 2 \\ 4 \end{bmatrix}$$

Étape 2. Soustrayons u'_2 à u'_3 et $2u'_2$ à u'_4 : on trouve la suite de vecteurs

$$u_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 3 \end{bmatrix}, \quad u'_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \end{bmatrix}, \quad u''_3 = u'_3 - u'_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad u''_4 = u'_4 - 2u'_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

et les vecteurs u_1, u'_2, u''_4 sont échelonnés.

Soit V le sous-espace vectoriel de $\mathbb{R}^4$ engendré par u_1, u_2, u_3, u_4. D'après la proposition, V est aussi engendré par u_1, u'_2, u''_3, u''_4, c'est-à-dire par u_1, u'_2, u''_4 puisque $u''_3 = 0$. Les trois vecteurs u_1, u'_2, u''_4 étant échelonnés, ils sont indépendants : les vecteurs u_1, u'_2, u''_4 forment donc une base de V. Par conséquent, on a $\dim V = 3$.

Résumé des méthodes pour trouver une base

Soit V un sous-espace vectoriel.

- Si l'on connaît des vecteurs $u_1, u_2, \ldots, u_n$ qui engendrent V, alors en échelonnant ces vecteurs, on obtient une base de V. On trouvera au plus n vecteurs en échelons, donc la dimension de V est inférieure ou égale à n.
- Si V est l'ensemble des solutions d'un système linéaire homogène à n inconnues, on résout ce système par la méthode de Gauss : si le système est de rang r, les solutions s'expriment de manière unique comme combinaison de $n-r$ vecteurs indépendants qui forment donc une base de V : $\dim V = n-r$.

3.4 Propriétés de la dimension

Propriétés 1. *Soit V un sous-espace vectoriel de $\mathbb{K}^p$.*

a) *Si V est engendré par n vecteurs, alors $\dim V \leqslant n$.*

b) *Si $u_1, u_2, \ldots, u_q$ sont des vecteurs indépendants appartenant à V, alors $q \leqslant \dim V$.*

c) *Supposons que V est contenu dans un sous-espace vectoriel W de $\mathbb{K}^p$. Alors on a $\dim V \leqslant \dim W$ et de plus, on a $V = W$ si et seulement si $\dim V = \dim W$.*

Démonstration. Posons $d = \dim V$ et soit $e_1, e_2, \ldots, e_d$ une base de V.

(a) Si l'on échelonne une suite de n vecteurs qui engendrent V, on obtient une base formée d'au plus n vecteurs en échelons, donc la dimension de V est inférieure ou égale à n.

(b) Chaque vecteur u_i est combinaison des vecteurs indépendants $e_1, \ldots, e_d$. Si l'on avait $q > d$, alors d'après la proposition page 115, les vecteurs $u_1, \ldots, u_q$ ne seraient pas indépendants. Par conséquent, on a $q \leqslant d$.

(c) Les vecteurs $e_1, \ldots, e_d$ appartiennent à V donc à W, et ils sont indépendants. D'après la propriété (b), on a donc $d \leqslant \dim W$. Supposons $d = \dim W$ et montrons que tout vecteur $u \in W$ appartient à V : puisqu'on a aussi $V \subset W$, cela impliquera l'égalité $V = W$. Raisonnons par l'absurde en supposant qu'il existe un vecteur $w \in W$ tel que $w \notin V$. Alors les vecteurs $e_1, e_2, \ldots, e_d, w$ sont deux à deux différents, donc leur nombre est $1+d = 1+\dim W$. Puisqu'ils appartiennent au sous-espace W de dimension d, ils ne sont pas indépendants (d'après (b)). On en déduit (théorème page 105) que w est combinaison linéaire de $e_1, \ldots, e_d$, ce qui est contraire à l'hypothèse. ∎

Propriétés 2. *Soit V un sous-espace vectoriel de $\mathbb{K}^p$ de dimension d.*

a) *Si $u_1, u_2, \ldots, u_d$ sont des vecteurs indépendants appartenant à V, ils forment une base de V.*

b) *Si $u_1, u_2, \ldots, u_d$ sont des vecteurs qui engendrent V, ils forment une base de V.*

Démonstration. Soit V' le sous-espace vectoriel engendré par $u_1, u_2, \ldots, u_d$. Puisque les vecteurs u_i appartiennent à V, tout vecteur de V' est un vecteur de V, autrement dit on a l'inclusion $V' \subset V$. Si les vecteurs $u_1, u_2, \ldots, u_d$ sont indépendants, ils forment une base de V', donc on a $\dim V' = d$, c'est-à-dire $\dim V' = \dim V$. D'après la propriété 1(c), on en déduit $V = V'$. Donc $u_1, u_2, \ldots, u_d$ est une base de V.

Supposons maintenant que les vecteurs $u_1, u_2, \ldots, u_d$ engendrent V. S'ils ne sont pas indépendants, l'un d'eux, u_d par exemple, est combinaison linéaire des autres. Alors V est engendré

par les vecteurs $u_1, u_2, \ldots, u_{d-1}$. Par la propriété 1(a), on en déduit $\dim V \leqslant d-1$, ce qui est une contradiction. Ce raisonnement montre que les vecteurs $u_1, u_2, \ldots, u_d$ sont indépendants. Puisqu'ils engendrent V, ils forment une base de V. ■

Propriété 3. *Si V est le sous-espace vectoriel de $\mathbb{K}^p$ engendré par des vecteurs $u_1, u_2, \ldots, u_n$, la dimension de V est égale au rang de l'équation linéaire $x_1u_1 + x_2u_2 + \cdots + x_nu_n = 0$.*

Démonstration. Notons r le rang de l'équation. Si $r = n$, cette équation homogène a pour unique solution $x_1 = \cdots = x_n = 0$ (résultat 2 page 114), donc les vecteurs $u_1, u_2, \ldots, u_n$ sont indépendants et l'on a bien $\dim V = n = r$.

Supposons maintenant $1 \leqslant r < n$. Supposons que, par la méthode de Gauss, on commence par éliminer x_1 dans la deuxième équation du système. On obtient une équation linéaire $x_1\tilde{u_1} + x_2\tilde{u_2} + \cdots + x_n\tilde{u_n} = 0$, où l'on est passé d'un vecteur u_i quelconque au vecteur $\tilde{u_i}$ en faisant sur les coordonnées la même opération :

si $u = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$ est l'un des vecteurs u_i, le vecteur $\tilde{u}$ correspondant est de la forme $\begin{bmatrix} a_1 \\ a_2+ta_1 \\ a_3 \\ \vdots \\ a_p \end{bmatrix}$,

et le scalaire t, choisi pour que la deuxième coordonnée de $\tilde{u_1}$ soit nulle, est le même pour tous les vecteurs u_i. Si l'on applique cette opération à un vecteur quelconque de $\mathbb{K}^p$, cela définit une transformation $u \mapsto \tilde{u}$ de $\mathbb{K}^p$. Le transformé de αu est $\alpha\tilde{u}$, le transformé de $u + v$ est $\tilde{u} + \tilde{v}$ et le transformé du vecteur nul est le vecteur nul.

Pour simplifier, supposons que la méthode de Gauss appliquée à l'équation

$$x_1u_1+x_2u_2+\cdots+x_nu_n = 0$$

conduise au système linéaire équivalent

$$(S') \qquad \begin{cases} x_1 + a'_{12}x_2 + a'_{13}x_3 + \cdots + a'_{1r}x_r + a'_{1\,r+1}x_{r+1} + \cdots + a'_{1n}x_n = 0 \\ \qquad x_2 + a'_{23}x_3 + \cdots + a'_{2r}x_r + a'_{2\,r+1}x_{r+1} + \cdots + a'_{2n}x_n = 0 \\ \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \qquad\qquad x_r + a'_{r\,r+1}x_{r+1} + \cdots + a'_{rn}x_n = 0 \end{cases}$$

(le système étant homogène, les égalités de compatibilité sont satisfaites.) Ce système s'écrit aussi $x_1u'_1 + x_2u'_2 + \cdots + x_nu'_n = 0$, où

$$u'_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad u'_2 = \begin{bmatrix} a'_{12} \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad u'_r = \begin{bmatrix} a'_{1r} \\ a'_{2r} \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad u'_{r+1} = \begin{bmatrix} a'_{1\,r+1} \\ a'_{2\,r+1} \\ \vdots \\ a'_{r\,r+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad u'_n = \begin{bmatrix} a'_{1n} \\ a'_{2n} \\ \vdots \\ a'_{rn} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

On passe des u_i aux u'_i par des opérations du même type que pour la transformation $u \mapsto \tilde{u}$; précisément, on effectue la même suite d'opérations sur tous les vecteurs, une opération élémentaire étant l'une des suivantes :

i) changer l'ordre des coordonnées,

ii) multiplier une coordonnée par un scalaire non nul,

iii) ajouter à l'une des coordonnées un multiple d'une autre coordonnée.

Pour tout vecteur $u \in \mathbb{K}^p$, notons u' le vecteur obtenu après cette suite d'opérations : cela définit une transformation $u \mapsto u'$ de $\mathbb{K}^p$. Le transformé de au est au', le transformé de $u+v$ est $u'+v'$ et le transformé du vecteur nul est le vecteur nul. De plus, on peut retrouver u à partir de u' par une suite d'opérations du même type, donc la transformation $u \mapsto u'$ est bijective. Montrons que les vecteurs $u_1, u_2, \ldots, u_r$ sont indépendants. Supposons qu'on a une relation $y_1u_1 + y_2u_2 + \cdots + y_ru_r = 0$; alors en transformant chaque membre de l'égalité, il vient $y_1u'_1 + y_2u'_2 + \cdots + y_ru'_r = 0$; les vecteurs $u'_1, u'_2, \ldots, u'_r$ étant visiblement indépendants (ils sont échelonnés par le bas), les scalaires $y_1, y_2, \ldots, y_r$ sont tous nuls.

Montrons que pour $r+1 \leqslant k \leqslant n$, le vecteur u_k est combinaison linéaire de $u_1, u_2, \ldots, u_r$. Le système d'équations aux inconnues $t_1, t_2, \ldots, t_r$:

$$\begin{cases} t_1 + a'_{12}t_2 + a'_{13}t_3 + \cdots + a'_{1r}t_r = a'_{1k} \\ \qquad t_2 + a'_{23}t_3 + \cdots + a'_{2r}t_r = a'_{2k} \\ \qquad\qquad \ddots \qquad \vdots \qquad \vdots \\ \qquad\qquad\qquad\qquad t_r = a'_{rk} \end{cases}$$

possède r équations, r inconnues et est de rang r, donc il a une unique solution $(t_1,t_2,\ldots,t_r)$. On a, dans $\mathbb{K}^r$, l'égalité de vecteurs

$$t_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + t_2 \begin{bmatrix} a'_{12} \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + t_r \begin{bmatrix} a'_{1r} \\ a'_{2r} \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} a'_{1k} \\ a'_{2k} \\ \vdots \\ a'_{rk} \end{bmatrix}$$

et comme les $p-r$ dernières coordonnées des vecteurs u'_i sont nulles, on a aussi

$$t_1u'_1 + t_2u'_2 + \cdots + t_ru'_r = u'_k\,.$$

Puisque la transformation $u \mapsto u'$ est une bijection, on en déduit $t_1u_1 + t_2u_2 + \cdots + t_ru_r = u_k$: chaque u_k est donc combinaison de $u_1, \ldots, u_r$. Ainsi le sous-espace vectoriel V engendré par $u_1, u_2, \ldots, u_n$ est aussi engendré par les vecteurs indépendants $u_1, \ldots, u_r$, donc V est de dimension r. ■

4. Un exemple d'application

Dans l'un de ses secteurs de production, une entreprise fabrique quatre sortes de fils synthétiques f_1, f_2, f_3, f_4, essentiellement à partir de deux matières premières M et M'. Dans le tableau ci-contre, les chiffres indiquent le pourcentage de M et le pourcentage de M' rentrant dans la composition de chaque fil.

	f_1	f_2	f_3	f_4
M	40	30	10	20
M'	10	30	10	50

Notons q_i le nombre d'unités de fil f_i produit chaque jour. Une production journalière est ainsi caractérisée par le vecteur (q_1,q_2,q_3,q_4). Cherchons les vecteurs de production qu'on peut obtenir à partir de quantités x et y des matières premières M et M'.

Les équations. Le fil f_1 utilise 40% de M et 10% de M', donc le nombre d'unités de fil f_1 produite est $q_1 = (4/10)x + (1/10)y$. En raisonnant de même pour les autres fils, on obtient les égalités :

$$(S) \qquad \begin{cases} (4/10)x + (1/10)y = q_1 \\ (3/10)x + (3/10)y = q_2 \\ (1/10)x + (1/10)y = q_3 \\ (2/10)x + (5/10)y = q_4 \end{cases}$$

Ce système d'équations est équivalent à :

$$\begin{cases} (4/10)x + (1/10)y = q_1 \\ (9/10)y = 4q_2 - 3q_1 \\ (3/10)y = 4q_3 - q_1 \\ (9/10)y = 2q_4 - q_1 \end{cases} \iff \begin{cases} (4/10)x + (1/10)y = q_1 \\ (9/10)y = 4q_2 - 3q_1 \\ 0 = 3(4q_3 - q_1) - (4q_2 - 3q_1) \\ 0 = (2q_4 - q_1) - (4q_2 - 3q_1) \end{cases}$$

$$\iff \begin{cases} (4/10)x + (1/10)1y = q_1 \\ (9/10)y = 4q_2 - 3q_1 \\ 0 = 3q_3 - q_2 \\ 0 = 2q_1 - 4q_2 + 2q_4 \end{cases} \iff \begin{cases} (4/10)x = q_1 - (4/9)q_2 + (1/3)q_1 \\ (9/10)y = 4q_2 - 3q_1 \\ 0 = 3q_3 - q_2 \\ 0 = q_1 - 2q_2 + q_4 \end{cases}$$

c'est-à-dire à : $(S') \quad \begin{cases} (1/10)x = (3q_1 - q_2)/9 \\ (9/10)y = 4q_2 - 3q_1 \\ 0 = 3q_3 - q_2 \\ 0 = q_1 - 2q_2 + q_4 \end{cases}$

Les quantités x, y, q_1, q_2, q_3 et q_4 devant être positives ou nulles, les conditions de production sont :

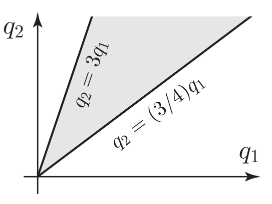
$$(C) \quad \begin{cases} 3q_1 \geqslant q_2 \geqslant 0 \\ 4q_2 \geqslant 3q_1 \geqslant 0 \end{cases} \qquad (C') \quad \begin{cases} 3q_3 = q_2 \\ q_4 = 2q_2 - q_1 \end{cases} \quad \text{et} \quad q_4 \geqslant 0.$$

D'après (C'), q_1 et q_2 déterminent q_3 et q_4. Si les conditions (C) sont satisfaites, alors $2q_2 \geqslant (3/2)q_1$ et l'on a bien $q_4 = 2q_2 - q_1 \geqslant (3/2)q_1 - q_1 = (1/2)q_1 \geqslant 0$.

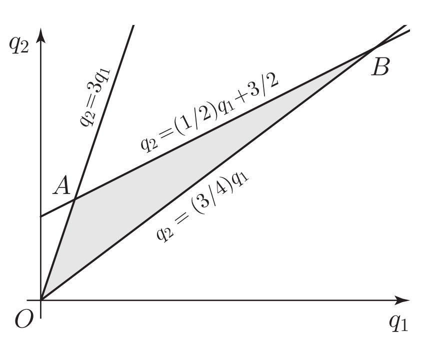
Un vecteur de production est donc déterminé par q_1 et q_2, pourvu que ces quantités vérifient (C). Les variables q_1 et q_2 sont des variables de contrôle. Pour des valeurs données de q_1 et q_2, les quantités x et y de matières premières nécessaires à la production se calculent au moyen des deux premières égalités de (S').

- Le vecteur de production $(q_1, q_2, q_3, q_4) = (3, 3, 1, 4)$ est irréalisable, car la seconde égalité de (C') n'est pas satisfaite.
- Le vecteur $(q_1, q_2, q_3, q_4) = (3,3,1,3)$ s'obtient avec les quantités de matières premières $x = 20/3$ et $y = 10/3$.

Représentation graphique. Sur le graphique ci-contre, on a porté q_1 en abscisse, q_2 en ordonnée, et l'on a représenté la droite d'équation $q_2 - 3q_1 = 0$ et la droite d'équation $4q_2 - 3q_1 = 0$. La partie grisée représente les couples (q_1, q_2) possibles.



Une contrainte supplémentaire. Pour des raisons commerciales, l'entreprise veut limiter à trois unités la production du fil f_4. On doit donc avoir $2q_2 - q_1 = q_4 \leqslant 3$, c'est-à-dire $q_2 \leqslant (1/2)q_1 + 3/2$. En ajoutant au graphique la droite d'équation $2q_2 - q_1 = 3$, on voit que les couples (q_1, q_2) possibles sont maintenant dans le triangle OAB représenté ci-contre. Les coordonnées du point A sont déterminées par les équations $q_2 = 3q_1 = (1/2)q_1 + 3/2$, d'où $(5/2)q_1 = 3/2$, $q_1 = 3/5$ et $q_2 = 9/5$. Les coordonnées du point B sont données par $q_2 = (3/4)q_1 = (1/2)q_1 + 3/2$, donc $(1/4)q_1 = 3/2$, $q_1 = 6$ et $q_2 = 9/2$.



$$A : (3/5, 9/5) \quad , \quad B : (6, 9/2)$$

Un problème d'optimisation. On demande en plus à l'entreprise de fournir un fil composite f constitué avec f_2, f_3 et f_4 et comportant 50% de f_4. Il est décidé de consacrer à cette production la moitié de la quantité q_3 et un tiers de q_4. Le bénéfice attendu concerne principalement la quantité du fil f_2 rentrant dans la nouvelle fabrication : l'entreprise cherche donc à rendre maximum cette quantité. Comment doit-elle ajuster les variables de contrôle q_1 et q_2 ?

Notons q la quantité de fil f_2 utilisée dans la nouvelle fabrication. On doit avoir l'égalité $\frac{q_4}{3} = \frac{q_3}{2} + q$ donc il s'agit de rendre maximum la quantité

$$q = \frac{q_4}{3} - \frac{q_3}{2} \, .$$

Exprimons q en fonction des variables de contrôle q_1 et q_2. D'après (C'), il vient $2q_4 - 3q_3 = 2(2q_2 - q_1) - q_2 = 3q_2 - 2q_1$, donc

$$(1) \qquad 6q = 3q_2 - 2q_1 \, .$$

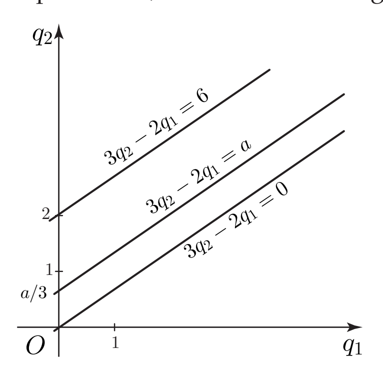
Rappelons que les variables q_1 et q_2 satisfont aux inégalités :

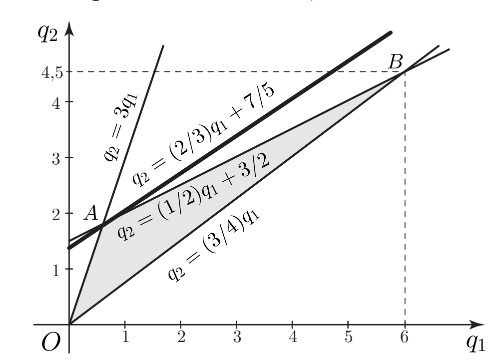
$$(2) \qquad \begin{cases} (3/4)q_1 \leqslant q_2 \leqslant 3q_1 \\ q_2 \leqslant (1/2)q_1 + 3/2 \end{cases}$$

qui expriment que le point de coordonnées (q_1, q_2) est dans le triangle OAB.

Dans le repère des variables (q_1, q_2), dessinons les droites d'équation $3q_2 - 2q_1 = a$, où a est un nombre quelconque. Ces droites ayant toutes la même pente $2/3$, elles

sont parallèles ; l'ordonnée à l'origine (obtenue pour $q_1 = 0$) est $a/3$.





La droite d'équation $3q_2 - 2q_1 = 6q$ est l'une de ces droites et elle doit contenir des points du triangle OAB. La pente $2/3$ étant comprise entre la pente de AB (qui vaut $1/2$) et celle de OA (qui vaut 3), c'est pour la droite passant par A que q est maximum. Cherchons l'équation de cette droite. Les coordonnées de A sont $(3/5, 9/5)$, donc $6q = 3(9/5) - 2(3/5) = 7(3/5)$ et $q = 7/10$. L'équation de la droite optimum est donc $3q_2 - 2q_1 = 6(7/10)$, ou encore $q_2 = (2/3)q_1 + 7/5$. Pour assurer cette production, il faut que $q_1 = 3/5$ et que $q_2 = 9/5$. Il vient alors $q_3 = q_2/3 = 3/5$ et $q_4 = 2q_2 - q_1 = 3$. Les quantités utilisées pour fabriquer le fil f sont

- pour f_4 : $q_4/3 = 1$,
- pour f_3 : $q_3/2 = 3/10$,
- pour f_2 : $q = 7/10$. La proportion de fil f_2 consacrée à la production de f est $\dfrac{q}{q_2} = \dfrac{7}{10}\dfrac{5}{9} = \dfrac{7}{18}$, soit un peu plus de $38,8\%$.

Exercices

1. On considère le système d'équations du pont de Wheatstone, page 104.

a) Montrer que l'intensité i_5 est nulle si et seulement si $r_1 r_3 = r_2 r_4$.

b) On suppose $r_1 = r_3 = r_4 = 2\Omega$ et $r_2 = 1\Omega$. Calculer les intensités (en ampère) en fonction de e (exprimé en volt) et de r_5. On suppose $e = 12V$ et l'on fait varier r_5 d'une valeur très petite à une valeur très grande. Entre quelles bornes varie l'intensité dans le générateur ?

@ **2.** On cherche une courbe polynomiale $x \mapsto P(x)$ de degré au plus 3 passant au point $A_1 = (x_1, y_1)$ avec une pente p_1 et au point $A_2 = (x_2, y_2)$ avec une pente p_2 (on suppose $x_1 \neq x_2$). Cherchons le polynôme P sous la forme $P(x) = a\left[(x-x_1)^2(x-x_2) + (x-x_1)(x-x_2)^2\right] + b(x-x_1)(x-x_2) + c(x-x_1) + d(x-x_2)$, où a, b, c, d sont des coefficients à calculer.

a) Écrire les quatre équations vérifiées par a, b, c, d.

b) Exprimer a, b, c, d au moyen de $x_1, x_2, y_1, y_2, p_1, p_2$.

c) On suppose $x_1 = y_1 = 0$, $p_1 = 5$, $x_2 = 3$, $y_2 = 2$ et $p_2 = 0$. Dessiner la courbe d'équation $y = P(x)$ sur l'intervalle $[x_1, x_2]$.

3. Résoudre le système d'équations linéaires $\begin{cases} 2x + y + z - t = 1 \\ 3x - y - z + t = 1 \end{cases}$

4. Résoudre le système d'équations linéaires $\begin{cases} x_1 - x_2 + x_3 - x_4 = a \\ x_1 + x_2 + x_3 + x_4 = b \\ x_1 + 2x_2 + x_3 + 2x_4 = c \end{cases}$

@ **5.** On considère le système d'équations linéaires $\begin{cases} x_1 + x_3 + 4x_4 + 4x_5 = a \\ x_1 + x_2 + 2x_3 + 2x_4 + 2x_5 = b \\ x_1 + 2x_2 + mx_3 = c \end{cases}$

a) Quel est le rang du système ? (discuter selon la valeur de m)

b) Pour quels nombres m ce système possède-t-il des solutions quels que soient a,b,c ?

c) Résoudre le système en discutant selon les valeurs de a, b, c, m.

@ **6.** Soit V le sous-espace vectoriel de $\mathbb{R}^4$ défini par les équations (S) $\begin{cases} y + z + t = 0 \\ x - y + z - t = 0 \\ x - 2y - 2t = 0 \\ x + 2z = 0 \end{cases}$

a) Trouver un système d'équations définissant V et comportant le nombre minimum d'équations. Quel est le rang du système (S) ?

b) Trouver une base de V. Quelle est la dimension de V ?

@ **7.** Trouver une base du sous-espace vectoriel de $\mathbb{R}^4$ engendré par les vecteurs $(0,1,2,1)$, $(1,-1,1,-1)$, $(a,0,3,0)$, $(1,1,5,1)$ (discuter selon les valeurs de a).

@ **8.** Posons $u = (0,1,2,3)$ et $v = (3,2,1,0)$. Soit V l'ensemble des vecteurs $(x,y,z,t) \in \mathbb{R}^4$ de la forme $au + bv$, où a et b parcourent les nombres réels.

a) Montrer que V est un sous-espace vectoriel de $\mathbb{R}^4$. Quelle est sa dimension ?

b) Trouver des équations de V.

@ **9.** Soit D la droite affine de $\mathbb{R}^3$ d'équations $\begin{cases} x - y + 2z = 1 \\ x + y + z = 1 \end{cases}$. Trouver l'équation du plan affine contenant D et passant par le point de coordonnées (a, a, a).

@ **10.** Soit V le plan vectoriel de $\mathbb{R}^3$ ayant pour base $(1,1,1), (1,0,2)$. Quelle est l'équation du plan affine passant par le point $A = (1, 2, 1)$ et de direction V ?

@11. Un exemple de systèmes à solutions positives. Considérons un système de n équations linéaires à n inconnues, de la forme

$$\left\{\begin{array}{l} a_{11}x_1 - a_{12}x_2 - a_{13}x_3 - \quad \cdots \quad - a_{1n}x_n = b_1 \\ - a_{21}x_1 + a_{22}x_2 - a_{23}x_3 - \quad \cdots \quad - a_{2n}x_n = b_2 \\ \quad \vdots \qquad \vdots \qquad\qquad\qquad\qquad \vdots \\ - a_{n1}x_1 - a_{n2}x_2 - \quad \cdots \quad - a_{n\,n-1}x_{n-1} + a_{nn}x_n = b_n \end{array}\right.$$

Faisons les deux hypothèses suivantes :

1) les nombres a_{ij} et les nombres b_i sont tous positifs ou nuls (ainsi les coefficients diagonaux sont positifs, les autres sont négatifs et le second membre est positif) ;

2) pour tout $i = 1, 2, \ldots n$, on a $a_{ii} > \sum_{j \neq i} a_{ij}$.

Un tel système est *à diagonale strictement dominante*. Le but de l'exercice est de montrer que si $(x_1, x_2, \ldots, x_n)$ est solution, les nombres x_i sont tous positifs ou nuls. Nous verrons aussi page 250 qu'un système vérifiant **(1)** et **(2)** possède une unique solution.

a) On suppose $n = 2$. Montrer qu'il y a une unique solution (x_1, x_2) et que l'on a $x_1 \geqslant 0$ et $x_2 \geqslant 0$.

b) On suppose $n = 3$. Soit (x_1, x_2, x_3) une solution. Supposons par exemple que le plus grand des nombres $|x_1|, |x_2|, |x_3|$ est $|x_2|$, donc on a $|x_2| \geqslant |x_1|$ et $|x_2| \geqslant |x_3|$.
(i) Montrer que $|a_{21}x_1 + a_{23}x_3| \leqslant a_{22}|x_2|$; en déduire que $-a_{21}x_1 + a_{22}x_2 - a_{23}x_3$ a le signe de $a_{22}x_2$ et que x_2 est positif ou nul.
(ii) Montrer que l'on a $\left\{\begin{array}{l} a_{11}x_1 - a_{13}x_3 \geqslant 0 \\ - a_{31}x_1 + a_{33}x_3 \geqslant 0 \end{array}\right.$. En utilisant le résultat **(a)**, en déduire que x_1 et x_3 sont positifs ou nuls.
(iii) Expliquer pourquoi le résultat est vrai quelle que soit la disposition des nombres $|x_1|, |x_2|, |x_3|$.

c) Expliquer le raisonnement par récurrence qui permet de démontrer le résultat pour un système de n équations.

Chapitre 5

Matrices et déterminants

1. Matrices

Rappelons que $\mathbb{K}$ désigne l'ensemble des nombres réels ou l'ensemble des nombres complexes.

1.1 Définitions et opérations sur les matrices

Définition

Un tableau formé de nombres appartenant à $\mathbb{K}$ s'appelle une *matrice* à coefficients dans $\mathbb{K}$.

Voici une matrice A à deux lignes et trois colonnes et une matrice B à deux lignes et deux colonnes :

$$A = \begin{bmatrix} a & ax & a^2 \\ by & b & y \end{bmatrix} \qquad B = \begin{bmatrix} u & t \\ 1 & 1/u \end{bmatrix}$$

L'ensemble des matrices à p lignes et n colonnes se note $\mathscr{M}_{p,n}(\mathbb{K})$. Une matrice ayant autant de lignes que de colonnes est dite *carrée* ; l'ensemble des matrices carrées à n lignes et n colonnes se note simplement $\mathscr{M}_n(\mathbb{K})$.

Une *matrice-ligne* est une matrice à une seule ligne, comme $[\,a_1 \;\; a_2 \;\; \cdots \;\; a_n\,]$; l'ensemble des matrices-ligne à n colonnes est donc $\mathscr{M}_{1,n}(\mathbb{K})$.

Nous avons déjà rencontré les *matrices-colonne* $\begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$ à p lignes : ce sont les éléments de $\mathscr{M}_{p,1}(\mathbb{K})$.

Si $u = (a_1, a_2, \dots, a_p)$ est un vecteur de $\mathbb{K}^p$, alors en disposant les coordonnées de u en colonne, on obtient la matrice-colonne $\begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$; en disposant les coordonnées en ligne, on obtient la matrice-ligne $[\,a_1 \;\; \cdots \;\; a_p\,]$. C'est pourquoi on dit aussi « vecteur-colonne » au lieu de « matrice-colonne », et « vecteur-ligne » au lieu de « matrice-ligne ».

Pour définir une matrice générale à p lignes et n colonnes, on utilise deux indices pour les coefficients :

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{bmatrix}$$

Le premier indice est le numéro de la ligne, le second est le numéro de la colonne. La matrice ci-dessus se note aussi $A = [a_{ij}]$.

Lignes et colonnes d'une matrice

- Si j est l'indice d'une colonne, la matrice-colonne $A_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix}$ s'appelle la j-ème colonne de A. Pour mettre en évidence les colonnes de A, on note $A = [\, A_1 \;\; A_2 \;\; \cdots \;\; A_n \,]$.
- De même, si i est l'indice d'une ligne, la matrice-ligne $L_i = [\, a_{i1} \;\; a_{i2} \;\; \cdots \;\; a_{in} \,]$ s'appelle la i-ème ligne de A et l'on pourra noter $A = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_p \end{bmatrix}$.
- Une matrice $[a]$ qui n'a qu'une seule ligne et qu'une seule colonne est identifiée au scalaire a.

Définition

Si $A = [a_{ij}]$ est une matrice carrée, les coefficients a_{ii} s'appellent les *coefficients diagonaux* de A. Une matrice carrée dont tous les coefficients non diagonaux sont nuls s'appelle une *matrice diagonale*. On note $\mathrm{diag}(a_1, a_2, \ldots, a_n)$ la matrice diagonale dont les coefficients diagonaux sont $a_1, \ldots, a_n$.

Ainsi par exemple, on a $\mathrm{diag}(a, b, c) = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$.

Produit d'une matrice par un scalaire

Pour multiplier une matrice A par un nombre t, on multiplie tous les coefficients par t : si $A = [a_{ij}]$, on a donc $tA = [ta_{ij}]$.

Somme de deux matrices

Soient $A = [a_{ij}]$ et $B = [b_{ij}]$ des matrices de même taille. On définit la matrice $A + B$ en ajoutant coefficient par coefficient, autrement dit le coefficient d'indices i, j dans la matrice $A + B$ est $a_{ij} + b_{ij}$. La matrice $A + B$ est de même taille que A et B.

Ces deux opérations permettent de définir des combinaisons linéaires de matrices de même taille.

Définition
Soient $A_1, A_2, \ldots, A_k$ des matrices à p lignes et n colonnes. Une *combinaison linéaire* des matrices $A_1, A_2, \ldots, A_k$ est une matrice de la forme
$$t_1A_1 + t_2A_2 + \cdots + t_kA_k\,,$$
où $t_1, t_2, \ldots, t_k$ sont des scalaires. C'est encore une matrice à p lignes et n colonnes.

Pour des matrices-colonnes à p lignes, on retrouve la notion de combinaison linéaire de vecteurs de $\mathbb{K}^p$.

Exemple. On a $2\begin{bmatrix} a & b & 0 \\ 0 & b' & c' \end{bmatrix} - 3\begin{bmatrix} 0 & v & w \\ u' & v' & 0 \end{bmatrix} = \begin{bmatrix} 2a & 2b-3v & -3w \\ -3u' & 2b'-3v' & 2c' \end{bmatrix}$.

Transposée d'une matrice

Définitions

- La *transposée* de la matrice-colonne $A = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$ est la matrice-ligne $\begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix}$ et se note ${}^tA$.
- Soit A une matrice ayant n colonnes $A_1, A_2, \ldots, A_n$. La *matrice transposée* de A, notée ${}^tA$, est la matrice dont les lignes sont ${}^tA_1, {}^tA_2, \ldots, {}^tA_n$.

Exemple. La transposée de la matrice $A = \begin{bmatrix} a & b & c \\ a' & b' & c' \end{bmatrix}$ est la matrice ${}^tA = \begin{bmatrix} a & a' \\ b & b' \\ c & c' \end{bmatrix}$.

Si $A \in \mathscr{M}_{p,n}(\mathbb{K})$, alors ${}^tA \in \mathscr{M}_{n,p}(\mathbb{K})$. Pour tous indices i et j, le coefficient situé en i-ème ligne, j-ème colonne de A se trouve en j-ème ligne, i-ème colonne de ${}^tA$. Si $A = [a_{ij}]$, on a donc ${}^tA = [a_{ji}]$ et l'égalité ${}^t({}^tA) = A$.

Si A est une matrice carrée, ${}^tA$ s'obtient en faisant une symétrie par rapport à la diagonale de A (c'est-à-dire la droite constituée des coefficients diagonaux).

Lorsqu'on a l'égalité $A = {}^tA$, on dit que la matrice A est *symétrique*.

Produit de matrices

Si A et B sont des matrices, le produit AB est défini seulement à la condition que le nombre de colonnes de A soit égal au nombre de lignes de B.

Procédons par étapes pour définir le produit matriciel AB. L'opération de base est le produit d'une ligne par une colonne.

A est une matrice-ligne et B est une matrice-colonne

Supposons $A = [\, a_1 \;\; a_2 \;\; \cdots \;\; a_n \,]$ et $B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ (ces matrices ayant le même nombre n de coefficients). On pose

$$AB = [\, a_1 \;\; a_2 \;\; \cdots \;\; a_n \,] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \,.$$

A est quelconque et B est une matrice-colonne

Supposons que $A = [a_{ij}]$ possède p lignes et n colonnes et que $B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ a n lignes. Pour tout indice i compris entre 1 et p, notons L_i la i-ème ligne de A. Le produit matriciel $L_i B$ est bien défini et vaut

$$L_i B = a_{i1} b_1 + a_{i2} b_2 + \cdots + a_{in} b_n \,.$$

On définit le produit AB de la manière suivante.

$$\text{La matrice produit } AB = \begin{bmatrix} L_1 \\ \vdots \\ L_p \end{bmatrix} B \text{ est la matrice-colonne } \begin{bmatrix} L_1 B \\ L_2 B \\ \vdots \\ L_p B \end{bmatrix} .$$

Le produit AB a autant de lignes que A.

A possède n colonnes et B possède n lignes

Appelons $C_1, C_2, \cdots, C_q$ les colonnes de B. Pour tout indice j compris entre 1 et q, le produit matriciel AC_j est bien défini : c'est une matrice-colonne ayant autant de lignes que A.

Définition

Le produit $AB = A[\, C_1 \;\; \cdots \;\; C_q \,]$ est la matrice $[\, AC_1 \;\; AC_2 \;\; \cdots \;\; AC_q \,]$. Le produit AB a autant de colonnes que B et autant de lignes que A.

Si $A \in \mathscr{M}_{p,n}(\mathbb{K})$ et si $B \in \mathscr{M}_{n,q}(\mathbb{K})$, alors le produit AB appartient à $\mathscr{M}_{p,q}(\mathbb{K})$.

Exemples

► Le produit $[\, a \;\; b \;\; c \,] \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = [a + 2b + 3c]$ s'identifie au scalaire $a + 2b + 3c \in \mathbb{K}$.

► Le produit $M=\begin{bmatrix} a & b & c \\ a' & b' & c' \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ est la matrice-colonne $\begin{bmatrix} m_{11} \\ m_{21} \end{bmatrix}$, avec $m_{11}=[a\ b\ c]\begin{bmatrix} x \\ y \\ z \end{bmatrix}=$ $ax+by+cz$ et $m_{21}=[a'\ b'\ c']\begin{bmatrix} x \\ y \\ z \end{bmatrix}=a'x+b'y+c'z$. On a donc $\begin{bmatrix} a & b & c \\ a' & b' & c' \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}=$ $\begin{bmatrix} ax+by+cz \\ a'x+b'y+c'z \end{bmatrix}$.

► Posons $A=\begin{bmatrix} 1 & 2 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}$, $B=\begin{bmatrix} x \\ y \end{bmatrix}$ et $C=\begin{bmatrix} x & z \\ y & t \end{bmatrix}$. Il vient

$$AB=\begin{bmatrix} x+2y \\ x-y \\ -x+y \end{bmatrix} \quad \text{et} \quad AC=\begin{bmatrix} 1 & 2 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} x & z \\ y & t \end{bmatrix}=\begin{bmatrix} x+2y & z+2t \\ x-y & z-t \\ -x+y & -z+t \end{bmatrix},$$

mais les produit BA et CA ne sont pas définis.

Produit de matrices carrées de même taille. Pour toutes matrices carrées A et B à n lignes et n colonnes, le produit AB est défini et c'est une matrice carrée à n lignes et n colonnes. On a par exemple $\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}\begin{bmatrix} b & c \\ 0 & b \end{bmatrix}=\begin{bmatrix} ab & ac+b \\ 0 & ab \end{bmatrix}$.
Voici une proposition simple et commode qui résulte immédiatement de la définition du produit matriciel.

Proposition. *Soit A une matrice à p lignes et n colonnes. Si les colonnes de A sont $A_1, A_2, \ldots, A_n$, alors pour tout vecteur-colonne $X=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, on a $AX = x_1A_1 + x_2A_2 + \cdots + x_nA_n$.*

Sélection d'une colonne ou d'une ligne. Rappelons que pour tout $j\in\{1,\ldots,n\}$, on note $\boldsymbol{e_j}$ le vecteur canonique de K^n dont tous les coefficients sont nuls sauf le j-ème qui vaut 1 et que $\boldsymbol{E_j}$ désigne le vecteur-colonne correspondant (voir page 102). D'après la proposition précédente,

si A est une matrice à p lignes et n colonnes, alors pour tout $j\in\{1,2,\ldots,n\}$, le produit $A\boldsymbol{E_j}$ est la j-ème colonne de A.

De même, si i est un indice entre 1 et p, la matrice ${}^t\boldsymbol{E_i}$ est une matrice-ligne et le produit $({}^t\boldsymbol{E_i})A$ est la i-ème ligne de A.
Dans une matrice, on peut donc sélectionner la colonne j en multipliant à droite par $\boldsymbol{E_j}$ et l'on peut sélectionner la ligne i en multipliant à gauche par ${}^t\boldsymbol{E_i}$.

Définition
Pour tout entier $n\geqslant 1$, on note $I_n=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & \cdots & 1 \end{bmatrix}$ la matrice diagonale $\mathrm{diag}(1,1,\ldots,1)$ à n lignes et n colonnes. Cette matrice s'appelle la *matrice unité* de $\mathscr{M}_n(\mathbb{K})$.

Corollaire. *Si A est une matrice à p lignes et n colonnes, alors on a $AI_n = A$ et $I_pA = A$.*

Démonstration. La j-ème colonne de I_n est $\boldsymbol{E_j} \in \mathscr{M}_{n,1}(\mathbb{K})$, donc par définition du produit de matrices, on a $AI_n = [\, A\boldsymbol{E_1} \;\; A\boldsymbol{E_2} \;\; \cdots \;\; A\boldsymbol{E_n} \,]$. Puisque $A\boldsymbol{E_j}$ est la j-ième colonne de A, les matrices AI_n et A ont les mêmes colonnes. On montre de même l'égalité $I_pA = A$ en remarquant que les lignes de I_p sont les matrices ${}^tE_i$. ■

1.2 Règles du calcul matriciel

On calcule évidemment sur les combinaisons linéaires de matrices comme sur les combinaisons linéaires de vecteurs. Concernant le produit de matrices, on a les règles suivantes.

I) $A(B + C) = AB + AC$ et $(A + B)C = AC + BC$.

II) Si $\lambda \in \mathbb{K}$, alors $A(\lambda B) = \lambda AB = (\lambda A)B$.

III) $A(BC) = (AB)C$ et cette matrice se note simplement ABC.

IV) ${}^t(AB) = ({}^tB)({}^tA)$.

- Il résulte des propriétés (I) et (II) que si $A \in \mathscr{M}_{p,n}(\mathbb{K})$ et si $B_1, B_2, \ldots, B_k$ sont des matrices de $\mathscr{M}_{n,q}(\mathbb{K})$, alors pour tous scalaires $x_1, x_2, \ldots, x_k$, on a dans $\mathscr{M}_{p,q}(\mathbb{K})$ l'égalité de matrices
$$A(x_1B_1 + x_2B_2 + \cdots + x_kB_k) = x_1AB_1 + x_2AB_2 + \cdots + x_kAB_k\,.$$
- La matrice de $\mathscr{M}_{p,n}(\mathbb{K})$ dont tous les coefficients sont nuls s'appelle la *matrice nulle* et se note simplement 0. Pour toute matrice $A \in \mathscr{M}_{p,n}(\mathbb{K})$, on a $A + 0 = 0 + A = A$ et si l'on multiplie une matrice par la matrice nulle, on obtient la matrice nulle.

Produit de matrices carrées. Le produit de deux matrices carrées de même taille est une matrice carrée de même taille : le produit de matrices est donc une opération dans l'ensemble $\mathscr{M}_n(\mathbb{K})$.
Si A est une matrice carrée de taille n, on peut définir son carré $AA = A^2$, son cube $AA^2 = A^3$ et de proche en proche, pour tout entier $k \geqslant 2$, sa puissance k-ème A^k, en posant $A^k = AA^{k-1}$. De plus, on pose $A^1 = A$ et $A^0 = I_n$.

Polynôme en une matrice. Soit $P = a_kz^k + a_{k-1}z^{k-1} + \cdots + a_1z + a_0$ un polynôme à coefficients dans $\mathbb{K}$. Pour toute matrice $A \in \mathscr{M}_n(\mathbb{K})$, on définit la matrice
$$P(A) = a_kA^k + a_{k-1}A^{k-1} + \cdots + a_1A + a_0I_n$$
qui appartient à $\mathscr{M}_n(\mathbb{K})$.

Particularités du produit matriciel

Nous avons vu que le produit de deux matrices n'est défini que si le nombre de colonnes de la première est égal au nombre de lignes de la seconde. Voici d'autres particularités du produit matriciel.

L'ordre des facteurs compte. Soient $A\in\mathscr{M}_{p,n}(\mathbb{K})$ et $B\in\mathscr{M}_{n,p}(\mathbb{K})$. Le produit AB est défini : c'est une matrice carrée de taille p ; et le produit BA est défini : c'est une matrice carrée de taille n. Si $p\neq n$, les matrices AB et BA ne sont évidemment pas égales. Mais même en supposant que A et B sont des matrices carrées ($p=n$), le produit AB n'est en général pas égal à BA : le produit des matrices n'est pas commutatif.

Exemple. Posons $A=\begin{bmatrix}1&1\\1&1\end{bmatrix}$ et $B=\begin{bmatrix}1&1\\-1&-1\end{bmatrix}$. On a

$$AB=\begin{bmatrix}1&1\\1&1\end{bmatrix}\begin{bmatrix}1&1\\-1&-1\end{bmatrix}=\begin{bmatrix}0&0\\0&0\end{bmatrix}\quad\text{et}\quad BA=\begin{bmatrix}1&1\\-1&-1\end{bmatrix}\begin{bmatrix}1&1\\1&1\end{bmatrix}=\begin{bmatrix}2&2\\-2&-2\end{bmatrix}.$$

En fin de paragraphe, nous mentionnons les principaux cas particuliers où l'ordre des facteurs ne compte pas.

Un produit de facteurs non nuls peut être nul. Autrement dit, si A et B sont des matrices, on peut avoir $A\neq 0$, $B\neq 0$ et $AB=0$. Par exemple, les matrices A et B de l'exemple précédent sont toutes deux non nulles et le produit AB est la matrice nulle. Comme autre exemple, prenons la matrice $N=\begin{bmatrix}0&1\\0&0\end{bmatrix}$: on a $N^2=\begin{bmatrix}0&1\\0&0\end{bmatrix}\begin{bmatrix}0&1\\0&0\end{bmatrix}=0$.

Dans une égalité, on ne peut pas toujours simplifier par un facteur non nul. C'est une conséquence de ce qui précède : supposons que A et B sont des matrices non nulles telles que $AB=0$, c'est-à-dire $AB=0B$; si l'on pouvait simplifier par B dans cette égalité, alors on aurait $A=0$, ce qui n'est pas vrai.

Exemple. Posons $A=\begin{bmatrix}1&1\\1&1\end{bmatrix}$ et $B=\begin{bmatrix}2&1\\-1&0\end{bmatrix}$. On a $AB=\begin{bmatrix}1&1\\1&1\end{bmatrix}\begin{bmatrix}2&1\\-1&0\end{bmatrix}=\begin{bmatrix}1&1\\1&1\end{bmatrix}$, donc $AB=A$. Cette égalité s'écrit $AB=AI_2$, mais B n'est pas égal à I_2 : on ne peut pas simplifier par A.

Nous verrons bientôt une condition permettant de faire ces simplifications.

Matrices qui commutent. Si des matrices carrées A et B de $\mathscr{M}_n(\mathbb{K})$ vérifient $AB=BA$, on dit qu'elles *commutent*. Voici des exemples de matrices qui commutent.

- Pour des matrices diagonales, on a l'égalité
$$\operatorname{diag}(a_1,\dots,a_n)\operatorname{diag}(b_1,\dots,b_n)=\operatorname{diag}(a_1b_1,\dots,a_nb_n)\,.$$
Il s'ensuit que si D et D' sont des matrices diagonales, alors $DD'=D'D$.
- Pour tout scalaire λ, la matrice $\lambda I_n=\operatorname{diag}(\lambda,\lambda,\dots,\lambda)$ commute avec toute autre matrice carrée A de taille n.

  En effet, puisque $AI_n=I_nA=A$, on a $A(\lambda I_n)=\lambda(AI_n)=\lambda A$ et $(\lambda I_n)A=\lambda(I_nA)=\lambda A$, d'après les règles de calcul.
- Une matrice $A\in\mathscr{M}_n(\mathbb{K})$ commute évidemment avec toutes ses puissances A^k, donc aussi avec une matrice $P(A)$, où P est un polynôme à coefficients dans $\mathbb{K}$. On en déduit que si P et Q sont des polynômes, alors les matrices $P(A)$ et $Q(A)$ commutent.

► Concernant la somme et le produit, les règles de calcul pour les polynômes en la matrice A sont les mêmes que pour les polynômes. On en déduit que pour tous polynômes P et Q, on a $P(A)Q(A) = (PQ)(A)$, où PQ est le polynôme produit. Ainsi par exemple, si $P = (2z-1)^k$ et si $A \in \mathscr{M}_n(\mathbb{K})$, alors $P(A) = (2A - I_n)^k$.

Pour des matrices qui commutent, on a aussi la formule du binôme comme pour les nombres.

Formule du binôme de Newton. *Soient A et B des matrices carrées de même taille. Si les matrices A et B commutent, alors pour tout entier $k \geqslant 2$, on a*

$$(A+B)^k = A^k + \binom{k}{1} A^{k-1}B + \cdots + \binom{k}{i} A^{k-i}B^i + \cdots + \binom{k}{k-1} AB^{k-1} + B^k$$

1.3 Matrices et systèmes linéaires

Considérons le système linéaire

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n = b_p \end{cases}$$

aux n inconnues $x_1, \ldots, x_n$. La *matrice du système* est par définition la matrice des coefficients du premier membre, c'est-à-dire

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{p1} & a_{22} & \cdots & a_{pn} \end{bmatrix}$$

Posons $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ et $B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$. Alors d'après la proposition page 133, le système s'écrit simplement

$$AX = B\,.$$

Formulons dans le cadre des matrices quelques-uns des résultats obtenus au chapitre précédent. Commençons par un résultat simple et fondamental.

Proposition. *Si X_0 est une solution du système linéaire $AX = B$, alors les solutions sont les vecteurs $X = X_0 + U$, où U est solution du système linéaire $AX = 0$.*

Démonstration. Supposons $AX_0 = B$. Pour tout vecteur-colonne $X \in \mathbb{K}^n$, on a les équivalences

$$AX = B \iff AX = AX_0 \iff AX - AX_0 = 0$$
$$\iff A(X - X_0) = 0 \iff X = X_0 + U\ ,\ \text{où } U \text{ vérifie } AU = 0. \qquad \blacksquare$$

Rappelons que les solutions du système linéaire $AX = 0$ forment un sous-espace vectoriel de $\mathbb{K}^n$ de dimension $n - r$, où r est le rang du système (page 116).

Définition
Soit A une matrice à p lignes et n colonnes. Le rang du système linéaire $AX = 0$ s'appelle le *rang de* A et se note $\operatorname{rg} A$.

Lorsqu'on résout un système linéaire par la méthode de Gauss, on pratique sur les coefficients des équations les opérations qui permettent d'échelonner les vecteurs-ligne de la matrice du système. À la fin, on obtient r vecteurs-ligne indépendants, où r est le rang du système. Puisque r vecteurs indépendants engendrent un sous-espace vectoriel de dimension r, on en déduit :

le rang d'une matrice est la dimension du sous-espace vectoriel engendré par les vecteurs-ligne.

D'après la propriété 3 page 121, on a aussi le théorème fondamental suivant.

Théorème. *Le rang d'une matrice à p lignes est la dimension du sous-espace vectoriel de $\mathbb{K}^p$ engendré par les vecteurs-colonne.*

Conséquence : *pour calculer le rang d'une matrice, on peut échelonner les vecteurs-ligne ou bien échelonner les vecteurs-colonne.*

1.4 Matrices inversibles

Proposition. *Soit A une matrice carrée de taille n. Pour toute matrice carrée C de taille n, on a l'équivalence $CA = I_n \iff AC = I_n$. De plus, s'il existe une telle matrice C, elle est unique.*

Démonstration. Supposons que C est une matrice carrée de taille n telle que $CA = I_n$. Si X est un vecteur-colonne de $\mathbb{K}^n$ tel que $AX = 0$, alors il vient $X = I_n X = (CA)X = C(AX) = C0 = 0$. L'équation $AX = 0$ a donc pour seule solution le vecteur nul. D'après le résultat 2 (b) page 114, cela veut dire que le rang de A est égal à n. Pour tout vecteur-colonne $U \in \mathbb{K}^n$, l'équation $AX = U$ a donc une unique solution, d'après le résultat 3 (c) page 114. Cette solution vérifie $X = (CA)X = C(AX) = CU$: l'unique solution de l'équation $AX = U$ est le vecteur CU. On a donc $A(CU) = U$, ou encore $(AC)U = U$, quel que soit le vecteur-colonne $U \in \mathbb{K}^n$.
Prenons pour U le i-ème vecteur canonique $\boldsymbol{E_i}$. On sait que le produit $(AC)\boldsymbol{E_i}$ est la i-ème colonne de AC (page 133). Puisque $(AC)\boldsymbol{E_i} = \boldsymbol{E_i}$, les colonnes de AC sont les $\boldsymbol{E_i}$, donc $AC = I_n$. Cela démontre l'implication $CA = I_n \Longrightarrow AC = I_n$.
Réciproquement, supposons que C est une matrice carrée de taille n telle que $AC = I_n$. En intervertissant les rôles de A et C dans le raisonnement précédent, on en déduit $CA = I_n$.
Il reste à montrer que si l'on a $CA = AC = I_n$, alors C est unique. Supposons que D est une (autre) matrice carrée vérifiant $DA = AD = I_n$. Alors il vient $D = DI_n = D(AC) = (DA)C = I_nC = C$. ■

Définition
Soit A une matrice carrée de taille n. On dit que A est *inversible* s'il existe une matrice carrée C telle que $CA = I_n$ (ou s'il existe une matrice carrée C telle que $AC = I_n$). La matrice C se note A^{-1} et s'appelle *l'inverse* de A.

Si $A \in \mathscr{M}_n(\mathbb{K})$ est une matrice inversible, on a donc les égalités $AA^{-1} = A^{-1}A = I_n$. Il s'ensuit que l'inverse de la matrice A^{-1} est A : on a $(A^{-1})^{-1} = A$.

Proposition. *Soit A une matrice carrée de taille n.*

i) *Si la matrice A est inversible, alors pour tout vecteur-colonne $U \in \mathbb{K}^n$, l'équation $AX = U$ a pour unique solution $X = A^{-1}U$.*

ii) *La matrice A est inversible si et seulement si* $\operatorname{rg} A = n$.

iii) *La matrice A est inversible si et seulement si l'équation $AX = 0$ a pour seule solution le vecteur nul.*

Démonstration. Supposons que A est inversible. Soient X et U des vecteurs-colonne de $\mathbb{K}^n$. Si $AX = U$, alors il vient $X = I_n X = (A^{-1}A)X = A^{-1}(AX) = A^{-1}U$. Réciproquement, si $X = A^{-1}U$, alors $AX = AA^{-1}U = I_n U = U$, d'où (i). Il s'ensuit que si A est inversible, alors l'équation $AX = 0$ a pour seule solution $X = A^{-1}0 = 0$. D'après les résultats du chapitre précédent, le rang de A est n.
Supposons réciproquement que $\operatorname{rg} A = n$. Alors pour tout vecteur colonne $U \in \mathbb{K}^n$, l'équation $AX = U$ a une unique solution (résultat 3 (c) page 114). Il y a donc des vecteurs-colonne C_i tels que $AC_i = \boldsymbol{E_i}$. Soit $C = [\,C_1 \ C_2 \ \cdots \ C_n\,]$ la matrice carrée de colonnes $C_1, C_2, \ldots, C_n$. Par définition du produit de matrices, les colonnes de AC sont $AC_1, AC_2, \ldots, AC_n$, c'est-à-dire $\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots, \boldsymbol{E_n}$: on a donc $AC = I_n$ et d'après la proposition précédente, la matrice A est inversible, d'inverse C. Cela montre (ii). Enfin, on sait que le rang de A est n si et seulement si l'équation $AX = 0$ a pour seule solution $X = 0$, d'où (iii). ■

Connaître l'inverse d'une matrice inversible A permet de résoudre <u>tous</u> les systèmes linéaires $AX = U$, quel que soit le vecteur-colonne U.

Propriétés des matrices inversibles

Soient A, B, C des matrices carrées de même taille.

i) Supposons A inversible. Si $AB = AC$, ou si $BA = CA$, alors $B = C$: on peut simplifier par une matrice inversible.

ii) Si A et B sont des matrices inversibles, alors la matrice produit AB est inversible et $(AB)^{-1} = B^{-1}A^{-1}$.

iii) Si A est inversible, alors ${}^tA$ est inversible et l'on a $({}^tA)^{-1} = {}^t(A^{-1})$.

Démonstration. Supposons A inversible et de taille n. Si $AB = AC$, alors en multipliant à gauche par la matrice A^{-1}, il vient $A^{-1}AB = A^{-1}AC$, c'est-à-dire $B = C$ puisque $A^{-1}A = I_n$. Si A et B sont inversibles, alors $(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}I_nB = B^{-1}B = I_n$, d'où (ii). On a aussi ${}^t(A^{-1}){}^tA = {}^t(AA^{-1}) = {}^tI_n = I_n$, ce qui montre (iii). ■

En particulier, si une matrice symétrique est inversible, son inverse est symétrique.

Exemples de matrices inversibles

Pour la matrice $A = \begin{bmatrix} a & p & q \\ 0 & b & r \\ 0 & 0 & c \end{bmatrix}$, l'équation linéaire $AX = 0$ s'écrit $\begin{cases} ax + py + qz = 0 \\ by + rz = 0 \\ cz = 0 \end{cases}$.

- Supposons que les coefficients diagonaux a, b, c sont tous différents de 0. La dernière équation donne $z = 0$; la deuxième s'écrit alors $by = 0$, donc $y = 0$ et la première s'écrit $ax = 0$, donc $x = 0$. On en déduit que la matrice A est inversible.
- Si par exemple $b = 0$, alors dans la résolution, peut choisir y quelconque : il y a donc d'autres solutions que le vecteur nul. Il s'ensuit que la matrice A n'est pas inversible. De même, si $c = 0$ ou si $a = 0$, la matrice n'est pas inversible.

Définition
Une matrice carrée est dite *triangulaire (supérieure)* si tous les coefficients situés strictement en-dessous de la diagonale sont égaux à 0.

Comme dans l'exemple ci-dessus, on démontre le résultat suivant.

Proposition. *Une matrice carrée triangulaire est inversible si et seulement si ses coefficients diagonaux sont tous différents de* 0. *L'inverse d'une matrice triangulaire supérieure (ou inférieure) est triangulaire supérieure (ou inférieure).*

Exemple. La matrice $A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & x & 0 & 0 \\ a & 4 & x & 0 \\ 0 & 3 & b & y \end{bmatrix}$ est inversible si et seulement si $xy \neq 0$.

D'après la proposition, une matrice diagonale $\text{diag}(a_1, a_2, \ldots, a_n)$ est inversible si et seulement si les a_i sont tous différents de 0 ; dans ce cas, on a

$$\left[\text{diag}(a_1, a_2, \ldots, a_n)\right]^{-1} = \text{diag}(1/a_1, 1/a_2, \ldots, 1/a_n)\,.$$

Calcul de l'inverse d'une matrice

Pour calculer l'inverse d'une matrice inversible A, on résout le système linéaire $A\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$, où les coefficients $u_1, \ldots, u_n$ sont des lettres : puisque la solution est $X = A^{-1}U$, les coefficients de A^{-1} sont ceux qui expriment les x_i au moyen de $u_1, \ldots, u_n$.

Exemple. Calculons l'inverse de la matrice $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 2 & 2 \end{bmatrix}$. On résout le système linéaire AX, où $U = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$:

$$\begin{cases} x = u \\ x + y = v \\ 3x + 2y + 2z = w \end{cases} \iff \begin{cases} x = u \\ y = -u + v \\ 2z = -3u - 2(-u+v) + w \end{cases} \iff \begin{cases} x = u \\ y = -u + v \\ 2z = -u - 2v + w \end{cases}$$

$$\iff \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1/2 & -1 & 1/2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

On a donc $A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1/2 & -1 & 1/2 \end{bmatrix}$.

Remarque

Dans les applications, où les matrices peuvent être de grande taille, on évite en général de calculer l'inverse (voir le chapitre 8 pour des méthodes numériques de résolution de systèmes).

1.5 Le groupe affine

Soit A une matrice carrée de taille n et soit B un vecteur-colonne de $\mathbb{R}^n$. La transformation de $\mathbb{R}^n$ définie par $X \mapsto AX + B$ s'appelle une *transformation affine*.

Exemple. L'application $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par $f(x, y) = (x + 2y + 1, 3x + 4y + 2)$ est la transformation affine de $\mathbb{R}^2$ telle que $\begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Propriétés des transformations affines

► *La composée de deux transformations affines de $\mathbb{R}^n$ est une transformation affine.*

Si f est la transformation $X \mapsto AX+B$ et si g est la transformation $X \mapsto CX+D$, la composée $g \circ f$ est définie par $(g \circ f)(X) = g(AX+B) = C(AX+B)+D = (CA)X+(CB+D)$.

► *La transformation affine $X \mapsto AX + B$ est bijective si et seulement si la matrice A est inversible. Dans ce cas, la transformation réciproque est $Y \mapsto A^{-1}Y - A^{-1}B$.*

En effet, pour tout vecteur $Y \in \mathbb{R}^n$, l'équation $AX + B = Y$ est équivalente à $AX = Y - B$. Pour que cette équation ait une unique solution quel que soit le vecteur Y, il faut et il suffit que la matrice A soit inversible ; dans ce cas, la solution est $X = A^{-1}(Y - B) = A^{-1}Y - A^{-1}B$.

En choisissant $A = I_n$ et $B = 0$, on obtient comme transformation affine l'application $X \mapsto X$, c'est-à-dire l'identité de $\mathbb{R}^n$.
Si les transformations affines f et g définies par $f(X) = AX + B$ et $g(X) = CX + D$ sont bijectives, la composée $g \circ f$ est bijective et définie par $X \mapsto (CA)X + (CB + D)$: la matrice CA, produit de deux matrices inversibles, est inversible.

Ces propriétés montrent que les transformations affines bijectives de $\mathbb{R}^n$ forment un groupe de transformations. Ce groupe s'appelle *le groupe affine.*

Transformations linéaires. Considérons seulement les transformations de $\mathbb{R}^n$ de la forme $X \mapsto AX$, où A est une matrice inversible. On dit que ce sont des transformations *linéaires*. La composée de deux transformations linéaires de $\mathbb{R}^n$ est une transformation linéaire et la transformation réciproque de $X \mapsto AX$ est $Y \mapsto A^{-1}Y$. Les transformations de la forme $X \mapsto AX$, où A est une matrice inversible, forment donc aussi un groupe de transformations de $\mathbb{R}^n$.

1.6 Exemple d'application : un intégrateur numérique

Le problème. Pour repérer le déplacement d'une pièce en mouvement sur un axe, on l'équipe d'un accéléromètre à inertie permettant de mesurer son accélération pendant l'intervalle de temps $[0, T]$. Les mesures ont lieu à intervalles réguliers, c'est-à-dire aux instants $t_1 = T/n$, $t_2 = 2T/n$, ..., $t_{n-1} = (n-1)T/n$, $t_n = T$. Soit γ_i l'accélération mesurée à l'instant t_i.
La position de la pièce sur son axe est déterminée à tout instant t par son abscisse $x(t)$. On dispose ainsi des nombres $x''(t_i) = \gamma_i$ et l'on cherche l'allure de la fonction $t \mapsto x(t)$. Il s'agit donc « d'intégrer deux fois » la fonction $t \mapsto x''(t)$ dont on ne connaît que les valeurs aux instants t_i.

Approximation de la dérivée seconde. Soit f une fonction deux fois dérivable sur un intervalle. Pour tout $t \in]a,b[$, on a la formule de Taylor-Young (page 301)

$$f(t+h) - f(t) = hf'(t) + \frac{h^2}{2} f''(t) + o(h^2)$$

où $o(h^2)$ désigne une quantité négligeable devant h^2 quand h tend vers 0. En remplaçant h par $-h$, il vient

$$f(t-h) - f(t) = -hf'(t) + \frac{h^2}{2} f''(t) + o(h^2)$$

et en ajoutant ces deux égalités, on obtient

$$f(t-h) - 2f(t) + f(t+h) = h^2 f''(t) + o(h^2)\,.$$

Ainsi le rapport $\dfrac{f(t-h) - 2f(t) + f(t+h)}{h^2}$ tend vers $f''(t)$ quand h tend vers 0. Pour h assez petit, le nombre $\frac{1}{h^2}\left[f(t-h) - 2f(t) + f(t+h)\right]$ est donc une bonne approximation de $f''(t)$.

Un procédé d'intégration numérique. Utilisons cette approximation pour évaluer les $x(t_i)$. Prenons $h = T/n$. Si n est assez grand, alors pour $i = 1, 2, \ldots, n-1$, le nombre

$$x(t_i - h) - 2x(t_i) + x(t_i + h) = x(t_{i-1}) - 2x(t_i) + x(t_{i+1})$$

est très proche de $h^2x''(t_i) = h^2\gamma_i$. On aura donc une bonne approximation des $x_i = x(t_i)$ en résolvant les égalités

$$(S) \qquad \begin{cases} x_0 - 2x_1 + x_2 = h^2\gamma_1 \\ x_1 - 2x_2 + x_3 = h^2\gamma_2 \\ x_2 - 2x_3 + x_4 = h^2\gamma_3 \\ \quad\vdots \qquad\qquad\quad \vdots \\ x_{n-2} - 2x_{n-1} + x_n = h^2\gamma_{n-1} \end{cases}$$

Détermination des solutions. Une fonction f n'est pas déterminée par sa dérivée seconde : en effet, si a et b sont des nombres quelconques, la fonction $t \mapsto g(t) = f(t)+at+b$ a pour dérivée seconde $g''=f''$. Pour déterminer g, on peut fixer sa valeur en deux points : si l'on connaît $g(u)$ et $g(v)$, où $u \neq v$, les égalités $au+b=g(u)-f(u)$ et $av+b=g(v)-f(v)$ permettent en effet de calculer les nombres a et b.
Dans notre cas, donnons-nous les valeurs $x_0 = x(0)$ et $x_n = x(T)$ aux bords de l'intervalle $[0, T]$. Pour $n = 6$, le système d'équations s'écrit

$$(S) \qquad \begin{cases} \quad - 2x_1 + x_2 = h^2\gamma_1 - x_0 \\ x_1 - 2x_2 + x_3 = h^2\gamma_2 \\ x_2 - 2x_3 + x_4 = h^2\gamma_3 \\ x_3 - 2x_4 + x_5 = h^2\gamma_4 \\ \quad x_4 - 2x_5 = h^2\gamma_5 - x_6 \end{cases}$$

Résolution. La matrice du système est $A = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}$. Ses coefficients non nuls sont concentrés sur la diagonale et de part et d'autre de celle-ci ; de plus, parallèlement à la diagonale, les coefficients sont égaux. Une telle matrice est dite *tridiagonale*. Calculons l'inverse de A en résolvant le système

$$\begin{cases} -2x_1 + x_2 = u_1 \\ x_1 - 2x_2 + x_3 = u_2 \\ x_2 - 2x_3 + x_4 = u_3 \\ x_3 - 2x_4 + x_5 = u_4 \\ x_4 - 2x_5 = u_5 \end{cases}$$

En multipliant par 5 la première équation, par 4 la deuxième, etc, et en ajoutant toutes les équations, les inconnues x_2, x_3, x_4, x_5 s'éliminent et l'on obtient simplement

$$-6x_1 = 5u_1 + 4u_2 + 3u_3 + 2u_4 + u_5\,.$$

De même, en multipliant la dernière équation par 5, l'avant-dernière par 4, etc, et en ajoutant, il vient

$$-6x_5 = u_1 + 2u_2 + 3u_3 + 4u_4 + 5u_5\,.$$

En reportant dans la première et dans la dernière équation, on en tire x_2 et x_4 :

$$-3x_2 = -3u_1 - 6x_1 = 2u_1 + 4u_2 + 3u_3 + 2u_4 + u_5$$
$$-3x_4 = -3u_5 - 6x_5 = u_1 + 2u_2 + 3u_3 + 4u_4 + 2u_5$$
$$-6x_3 = 3u_3 - 3x_2 - 3x_4 = 3u_1 + 6u_2 + 9u_3 + 6u_4 + 3u_5\,.$$

L'inverse de la matrice A est donc $A^{-1} = -\frac{1}{6}\begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 8 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 8 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$.

Un exemple numérique. On fait cinq mesures d'accélération, une toute les vingt secondes : ces mesures sont donc effectuées aux instants $t_i = 20i$, où $i = 1, \ldots, 5$. On a $h = 20$, $n = 6$ et $T = nh = 120$. Voici les valeurs mesurées (γ_i est en ms^{-2}) :

t_i	20	40	60	80	100
$\gamma_i \times 10^3$	2,44	1,20	−0,03	−1,27	−2,51

Si les positions initiales et finales (exprimées en mètre) sont $x(0) = x_0 = 0$ et $x(T) = x_6 = 1$, alors les valeurs approchées $x_1, x_2, \ldots, x_5$ sont données par

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = A^{-1}\begin{bmatrix} h^2\gamma_1 - 0 \\ h^2\gamma_2 \\ h^2\gamma_3 \\ h^2\gamma_4 \\ h^2\gamma_5 - 1 \end{bmatrix} = \begin{bmatrix} -0,63 \\ -0,27 \\ 0,56 \\ 1,38 \\ 1,69 \end{bmatrix}$$

La figure 1 ci-dessous montre la ligne polygonale des points (t_i, x_i) pour $i = 0, 2, \ldots, 6$. Sur la figure 2, on propose une courbe $t \mapsto x(t)$ deux fois dérivable ayant la même allure.

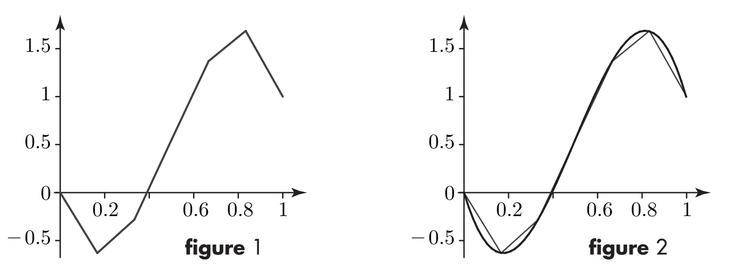


figure 1 **figure 2**

Cas général. Pour n quelconque, la matrice du système (S) est

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & \cdots & 1 & -2 & 1 \\ 0 & & \cdots & 0 & 1 & -2 \end{bmatrix}.$$

Cette matrice est symétrique et tridiagonale. On vérifie facilement que l'inverse de A est la matrice symétrique dont le coefficient en position i-ème ligne, j-ème colonne est $-\frac{1}{n+1}i(n+1-j)$ si $i \leqslant j$ et $-\frac{1}{n+1}j(n+1-i)$ si $i > j$.

2. Déterminants

Nous allons présenter la notion de déterminant d'une matrice carrée : c'est un outil théorique puissant, notamment pour caractériser les matrices inversibles.

2.1 Déterminant d'une matrice carrée de taille 1, 2 ou 3

Définitions

- Le déterminant d'une matrice $[a]$ de taille 1 est le scalaire a.
- Le déterminant de la matrice $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ est le nombre $ad - bc$. On le note $\det A$ ou $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$.

On a l'égalité $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad-bc & 0 \\ 0 & -bc+ad \end{bmatrix} = (ad-bc)I_2 = (\det A)I_2$. Il s'ensuit que si $\det A$ n'est pas nul, alors la matrice A est inversible et

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

On a $\det I_2 = 1$ et les propriétés immédiates suivantes.

i) Si les colonnes de la matrice A sont égales, alors le déterminant est nul.

ii) Le déterminant est additif par rapport à chaque vecteur-colonne :

$$\begin{vmatrix} a+a' & b \\ c+c' & d \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b \\ c' & d \end{vmatrix} \text{ et } \begin{vmatrix} a & b+b' \\ c & d+d' \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a & b' \\ c & d' \end{vmatrix}.$$

iii) Si l'on multiplie une colonne par un scalaire λ, alors le déterminant est multiplié par λ : en effet, on a $\begin{vmatrix} \lambda a & b \\ \lambda c & d \end{vmatrix} = \lambda \begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} a & \lambda b \\ c & \lambda d \end{vmatrix}$.

Définition

Le déterminant de la matrice $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ est le nombre

$$\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Dans cette expression, les coefficients de A en facteur des déterminants sont ceux de la première ligne : on dit qu'on a *développé le déterminant selon la première ligne*.

Pour avoir le déterminant de taille 2 associé dans la formule à un coefficient c de la première ligne,

► on supprime dans A la première ligne et la colonne de c

► et l'on prend le déterminant formé des quatre coefficients restants.

Noter l'alternance des signes.

Exemples

► On a

$$\begin{aligned}\begin{vmatrix} 1 & 2 & -1 \\ 2 & -1 & x \\ -1 & x & 1 \end{vmatrix} &= 1 \times \begin{vmatrix} -1 & x \\ x & 1 \end{vmatrix} - 2 \times \begin{vmatrix} 2 & x \\ -1 & 1 \end{vmatrix} - 1 \times \begin{vmatrix} 2 & -1 \\ -1 & x \end{vmatrix} \\ &= (-1 - x^2) - 2 \times (2 + x) - (2x - 1) \\ &= -1 - x^2 - 4 - 2x - 2x + 1 = -(x^2 + 4x + 4) = -(x + 2)^2 .\end{aligned}$$

► Pour une matrice triangulaire inférieure, il vient $\begin{vmatrix} a & 0 & 0 \\ x & b & 0 \\ y & z & c \end{vmatrix} = a \begin{vmatrix} b & 0 \\ z & c \end{vmatrix} = abc$, produit des coefficients diagonaux. En particulier, $\det I_3 = 1$.

Les déterminants de taille 3 possèdent encore les propriétés mises en évidence dans le cas de la taille 2. Notons A_1, A_2, A_3 les vecteurs-colonne de la matrice et $\det(A_1, A_2, A_3)$ le déterminant.

i) Si deux colonnes sont égales, alors le déterminant est nul : par exemple, si $A_1 = A_2$, alors le déterminant $\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$ est nul, on a $\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$ et comme $a_{11} = a_{12}$, il vient $\det A = 0$.

ii) Le déterminant est additif par rapport à chaque vecteur-colonne : cela veut dire que l'on a $\det(A_1 + A_1', A_2, A_3) = \det(A_1, A_2, A_3) + \det(A_1', A_2, A_3)$ et de même pour $\det(A_1, A_2 + A_2', A_3)$ et $\det(A_1, A_2, A_3 + A_3')$.

iii) Si l'on multiplie une colonne par un scalaire λ, alors le déterminant est multiplié par λ : $\det(\lambda A_1, A_2, A_3) = \det(A_1, \lambda A_2, A_3) = \det(A_1, A_2, \lambda A_3) = \lambda \det(A_1, A_2, A_3)$.

2.2 Déterminant d'une matrice carrée de taille n

On définit le déterminant d'une matrice carrée de taille n au moyen du déterminant des matrices de taille $n-1$, comme on l'a fait dans le cas $n = 3$.

Définition

Soit A une matrice carrée de taille n. Pour tous indices i et j, le *cofacteur* Δ_{ij} est le déterminant de taille $n-1$ obtenu en supprimant la i-ème ligne et la j-ème colonne de A.

► Pour la matrice $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, on a $\Delta_{11} = d$, $\Delta_{12} = c$, $\Delta_{21} = b$ et $\Delta_{22} = a$.

► Pour une matrice $A = [A_{ij}]$ de taille 3, on a $\Delta_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$, $\Delta_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$ et $\Delta_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$. Par définition du déterminant de A, il vient donc

$$\det A = a_{11}\Delta_{11} - a_{12}\Delta_{12} + a_{13}\Delta_{13}\,.$$

Définition

Le déterminant d'une matrice carrée $A = [a_{ij}]$ de taille n est le nombre

$$\det A = a_{11}\Delta_{11} - a_{12}\Delta_{12} + \cdots + (-1)^{1+j} a_{1j}\Delta_{1j} + \cdots + (-1)^{1+n} a_{1n}\Delta_{1n}\,.$$

Cette expression s'appelle *le développement du déterminant selon la première ligne.*

Propriétés relatives aux colonnes

En raisonnant par récurrence sur la taille du déterminant, on démontre en utilisant la définition que $\det I_n = 1$ et qu'un déterminant de taille n possède encore les trois propriétés précédentes. Formulons ces propriétés dans ce cadre général, en notant $A_1, A_2, \ldots, A_n$ les colonnes de la matrice.

Propriété 1. Si deux colonnes de la matrice sont égales, alors le déterminant est nul.

Propriété 2. Le déterminant est additif par rapport à chaque vecteur-colonne :

$$\det(A_1, \ldots, A_j + A'_j, \ldots, A_n) = \det(A_1, \ldots, A_j, \ldots, A_n) + \det(A_1, \ldots, A'_j, \ldots, A_n)\,.$$

Propriété 3. Si l'on multiplie une colonne par λ, le déterminant est multiplié par λ :

$$\det(A_1, \ldots, \lambda A_j, \ldots, A_n) = \lambda \det(A_1, \ldots, A_j, \ldots, A_n)\,.$$

Voici une conséquence de la propriété 3 :

Propriété 3*. Si une colonne est nulle, le déterminant est nul.

En effet, si l'on multiplie une colonne nulle par 0, elle ne change pas, mais le déterminant est multiplié par 0.

On en déduit aussi les propriétés suivantes.

Propriété 4. Si l'on permute deux colonnes de la matrice, le déterminant est changé de signe.

Considérons le déterminant $d = \det(A_1 + A_2, A_1 + A_2, A_3, \ldots, A_n)$. Puisque les deux premières colonnes sont égales, on a $d = 0$ (propriété 1). En utilisant la propriété 2, il vient

$$\begin{aligned} d &= \det(A_1, A_1 + A_2, A_3, \ldots, A_n) + \det(A_2, A_1 + A_2, A_3, \ldots, A_n) \\ &= \det(A_1, A_1, A_3, \ldots, A_n) + \det(A_1, A_2, A_3, \ldots, A_n) \\ &\qquad + \det(A_2, A_1, A_3, \ldots, A_n) + \det(A_2, A_2, A_3, \ldots, A_n) \\ &= 0 + \det(A_1, A_2, A_3, \ldots, A_n) + \det(A_2, A_1, A_3, \ldots, A_n) + 0 \qquad \text{d'après 1}\,, \end{aligned}$$

d'où $0 = \det(A_1, A_2, \ldots, A_n) + \det(A_2, A_1, \ldots, A_n)$.

Propriété 5.

► Si une colonne est combinaison linéaire des autres, le déterminant est nul.

► Si l'on ajoute à une colonne une combinaison linéaire des autres, le déterminant garde la même valeur.

En effet, si $U = \lambda_2 A_2 + \cdots + \lambda_n A_n$ est une combinaison linéaire de $A_2, \ldots, A_n$, on a $\det(U, A_2, \ldots, A_n) = \lambda_2 \det(A_2, A_2, \ldots, A_n) + \cdots + \lambda_n \det(A_n, A_2, \ldots, A_{n-1}, A_n) = 0$ car au second membre, chaque déterminant a deux colonnes égales. On en déduit $\det(A_1 + U, A_2, \ldots, A_n) = \det(A_1, A_2, \ldots, A_n) + \det(U, A_2, \ldots, A_n) = \det(A_1, A_2, \ldots, A_n)$.

Propriétés relatives aux lignes

Nous allons montrer que le déterminant possède, par rapport aux lignes, les mêmes propriétés que ci-dessus.

Propriété 1'. Si deux lignes de la matrice sont égales, alors le déterminant est nul.

Propriété 2'. Le déterminant est additif par rapport à chaque vecteur-ligne.

Propriété 3'. Si l'on multiplie une ligne par λ, le déterminant est multiplié par λ.

Démonstration. On vérifie immédiatement que ces propriétés sont vraies pour un déterminant de taille 2. Raisonnons par récurrence sur la taille du déterminant. Supposons que les propriétés 1', 2' et 3' sont vraies pour les déterminants de taille $n-1$ et considérons une matrice A de taille n, de vecteurs-lignes $L_1, L_2, \ldots, L_n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & \boxed{\quad L_2 \quad} & & \\ & \vdots & & \\ & \boxed{\quad L_n \quad} & & \end{bmatrix}$$

Si l'on multiplie la première ligne par λ, alors dans le développement selon la première ligne

$$\det A = a_{11}\Delta_{11} - a_{12}\Delta_{12} + \cdots + (-1)^{1+j}\Delta_{1j} + \cdots + (-1)^{1+n}\Delta_{1n} \, ,$$

les Δ_{1j} sont inchangés, donc le déterminant est multiplié par λ. Si l'on multiplie une autre ligne par λ, alors dans chaque Δ_{1j} une ligne est multiplié par λ, donc Δ_{1j} est multiplié par λ par hypothèse de récurrence : le déterminant de A est donc aussi multiplié par λ. En raisonnant comme pour les colonnes, il s'ensuit que si une ligne est nulle, alors le déterminant est nul. De même, il est clair que le déterminant est additif par rapport à la première ligne ; par hypothèse de récurrence, les Δ_{1j} sont additifs par rapport aux lignes, donc aussi le déterminant de A. Si deux lignes sont égales parmi les lignes numéro $2, 3, \ldots, n$, alors chaque Δ_{1j} possède deux lignes égales, donc est nul : il s'ensuit que le déterminant de A est nul. Supposons maintenant que la ligne L_1 est égale à la ligne L_k, où $k \geqslant 2$. Si cette ligne est nulle, le déterminant est nul d'après le développement selon la première ligne. Supposons que dans la ligne $L_1 = L_k$, le premier coefficient non nul est a_{1m}, donc

$$L_1 = [\,0 \ \cdots \ 0 \ a_{1m} \ \cdots \ a_{1n}\,] = L_k \, .$$

Pour tout $j \geqslant m+1$, remplaçons la colonne A_j par la combinaison $A'_j = A_j - (a_{1j}/a_{1m})A_m$, ce qui ne change pas le déterminant d'après la propriété 5. Sur les lignes 1 et k, le coefficient de A'_j est $a_{1j} - (a_{1j}/a_{1m})a_{1m} = 0$, donc

$$\det A = \det(A_1, \ldots, A_m, A'_{m+1}, \ldots, A'_n) = \begin{vmatrix} 0 & \cdots & 0 & a_{1m} & 0 & \cdots & 0 \\ & & & \boxed{\quad * \quad} & & & \\ 0 & \cdots & 0 & a_{1m} & 0 & \cdots & 0 \\ & & & \boxed{\quad * \quad} & & & \end{vmatrix}$$

Développons par rapport à la première ligne : il vient $\det A = (-1)^{1+m} a_{1m}\Delta_{1m}$ et dans le déterminant Δ_{1m}, la ligne numéro $k-1$ est nulle (elle est en position k dans A). Nous avons remarqué ci-dessus que cela implique $\Delta_{1m} = 0$, donc $\det A = 0$. ■

Comme pour les colonnes, on en déduit les propriétés suivantes.

Propriété 3'*. Si une ligne est nulle, le déterminant est nul.

Propriété 4'. Si l'on permute deux lignes de la matrice, le déterminant est changé de signe.

Propriété 5'. Si l'on ajoute à une ligne une combinaison linéaire des autres, le déterminant garde la même valeur.

2.3 Les théorèmes fondamentaux

Proposition. *Si f est une fonction de $\mathscr{M}_n(\mathbb{K})$ dans $\mathbb{K}$ possédant les propriétés 1, 2 et 3, alors pour toute matrice $M \in \mathscr{M}_n(\mathbb{K})$, on a $f(M) = f(I_n)\det({}^tM)$.*

Justification. Expliquons cela pour $n = 3$. Supposons que $f : \mathscr{M}_3(\mathbb{K}) \to \mathbb{K}$ est une fonction possédant les propriétés 1, 2 et 3, donc aussi la propriété 4. Pour $j = 1,2,3$, notons (a_{1j}, a_{2j}, a_{3j}) les coordonnées du vecteur-colonne A_j de la matrice A. On a $A_j = a_{1j}\boldsymbol{E_1} + a_{2j}\boldsymbol{E_2} + a_{3j}\boldsymbol{E_3}$, où les $\boldsymbol{E_i}$ sont les vecteurs canoniques de $\mathbb{K}^3$. D'après la propriété 2, il vient

$$\begin{aligned} f(\boldsymbol{E_1}, A_2, A_3) &= f(\boldsymbol{E_1}, A_2, a_{13}\boldsymbol{E_1}) + f(\boldsymbol{E_1}, A_2, a_{23}\boldsymbol{E_2}) + f(\boldsymbol{E_1}, A_2, a_{33}\boldsymbol{E_3}) \\ &= a_{13}f(\boldsymbol{E_1}, A_2, \boldsymbol{E_1}) + a_{23}f(\boldsymbol{E_1}, A_2, \boldsymbol{E_2}) + a_{33}f(\boldsymbol{E_1}, A_2, \boldsymbol{E_3}), \text{ grâce à 3,} \\ &= a_{23}f(\boldsymbol{E_1}, A_2, \boldsymbol{E_2}) + a_{33}f(\boldsymbol{E_1}, A_2, \boldsymbol{E_3}), \text{ car } f(\boldsymbol{E_1}, A_2, \boldsymbol{E_1}) = 0. \end{aligned}$$

$$\begin{aligned} f(\boldsymbol{E_1}, A_2, \boldsymbol{E_2}) &= a_{12}f(\boldsymbol{E_1}, \boldsymbol{E_1}, \boldsymbol{E_2}) + a_{22}f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_2}) + a_{32}f(\boldsymbol{E_1}, \boldsymbol{E_3}, \boldsymbol{E_2}) \\ &= a_{32}f(\boldsymbol{E_1}, \boldsymbol{E_3}, \boldsymbol{E_2}) = -a_{32}f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3}), \text{d'après 1 et 4.} \end{aligned}$$

De même, on a $f(\boldsymbol{E_1}, A_2, \boldsymbol{E_3}) = a_{22}f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3})$ d'où

$$f(\boldsymbol{E_1}, A_2, A_3) = (-a_{23}a_{32} + a_{33}a_{22})f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3})\,.$$

On obtient de la même manière

$$\begin{aligned} f(\boldsymbol{E_2}, A_2, A_3) &= (-a_{12}a_{33} + a_{13}a_{32})f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3}) \\ f(\boldsymbol{E_3}, A_2, A_3) &= (a_{12}a_{23} - a_{13}a_{22})f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3}) \end{aligned}$$

et il vient

$$\begin{aligned} f(A_1, A_2, A_3) &= f(a_{11}\boldsymbol{E_1} + a_{21}\boldsymbol{E_2} + a_{31}\boldsymbol{E_3}, A_2, A_3) \\ &= a_{11}f(\boldsymbol{E_1}, A_2, A_3) + a_{21}f(\boldsymbol{E_2}, A_2, A_3) + a_{31}f(\boldsymbol{E_3}, A_2, A_3)\ , \text{ d'où} \end{aligned}$$

$$f(A_1,A_2,A_3) = \big[a_{11}(a_{33}a_{22} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22})\big] f(\boldsymbol{E_1},\boldsymbol{E_2},\boldsymbol{E_3}).$$

Le crochet est égal à

$$a_{11}\begin{vmatrix} a_{22} & a_{32} \\ a_{23} & a_{33} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{32} \\ a_{13} & a_{33} \end{vmatrix} + a_{31}\begin{vmatrix} a_{12} & a_{22} \\ a_{13} & a_{23} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} = \det({}^tM)$$

par définition du développement de ${}^tM$ selon sa première ligne.

La fonction f est donc parfaitement déterminée par le nombre $f(\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3}) = f(I_3)$. Si l'on a $f(I_3) = 1$, alors $f(M) = \det({}^tM)$ quelle que soit M. Dans le cas général, posons $a = f(I_3)$. Si $a \neq 0$, alors la fonction $g : M \mapsto (1/a)f(M)$ possède les propriétés 1, 2 et 3 et $g(I_3) = (1/a)f(I_3) = 1$,

donc $g(M) = \det({}^tM)$ et dans ce cas, on a bien $f(M) = ag(M) = a\det({}^tM)$. Si $a = 0$, alors la fonction f est nulle et l'on a encore $f(M) = a\det({}^tM)$ pour tout M. ■

On en déduit de nouvelles et importantes propriétés du déterminant.

Proposition. *Pour toutes matrices carrées de taille n, on a $\det({}^tA) = \det A$ et $\det(AB) = (\det A)(\det B)$.*

Démonstration. Pour toute matrice $M \in \mathscr{M}_n(\mathbb{K})$, posons $f(M) = \det(AM)$. Par définition du produit de matrices, on a $f(M) = \det(AM_1, \ldots, AM_n)$, où $M_1, \ldots, M_n$ sont les vecteurs-colonne de M et cela montre que f possède les propriétés 1, 2 et 3. Puisque $f(I_n) = \det(AI_n) = \det A$, on déduit de la précédente proposition que l'on a $\det(AM) = f(M) = (\det A)(\det {}^tM)$ pour tout M. En particulier, pour $A = I_n$, il vient $\det M = (\det I_n)(\det {}^tM) = \det {}^tM$, ce qui est la première propriété. On en déduit alors l'égalité $\det(AM) = (\det A)(\det {}^tM) = (\det A)(\det M)$, pour tout M. ■

Montrons que le déterminant permet de caractériser les matrices inversibles.

Proposition. *Une matrice carrée est inversible si et seulement si son déterminant est non nul. Si A est une matrice inversible, alors $\det(A^{-1}) = \dfrac{1}{\det A}$.*

Démonstration. Soit $A \in \mathscr{M}_n(\mathbb{K})$. Si A est inversible, alors $(\det A)(\det A^{-1}) = \det(AA^{-1}) = \det I_n = 1$, donc le nombre $\det A$ n'est pas nul et $\det(A^{-1})$ est l'inverse de $\det A$. Supposons que la matrice A n'est pas inversible. Alors $\operatorname{rg} A$ est strictement inférieur à n, donc les vecteurs-colonne de A ne sont pas indépendants (proposition page 138 et théorème page 137). Il s'ensuit qu'au moins l'un des vecteurs-colonne est combinaison linéaire des autres, donc le déterminant de A est nul, d'après la propriété 5. ■

Exemple. La matrice $\begin{bmatrix} 1 & 2 & -1 \\ 2 & -1 & x \\ -1 & x & 1 \end{bmatrix}$ de l'exemple page 145 est inversible si et seulement si $x \neq -2$.

Développement selon une ligne ou une colonne

Le déterminant a été défini par son développement selon la première ligne. Mais échangeons par exemple la première et la deuxième ligne dans la matrice $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$. On obtient $A' = \begin{bmatrix} a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ et en développant le déterminant de A' selon la première ligne, il vient $\det A' = a_{21}\Delta_{21} - a_{22}\Delta_{22} + a_{23}\Delta_{23}$, où les Δ_{ij} sont les cofacteurs pour la matrice A. Puisque $\det A = -\det A'$, on a le développement

$$\det A = -a_{21}\Delta_{21} + a_{22}\Delta_{22} - a_{23}\Delta_{23}\,.$$

Plus généralement, on peut développer un déterminant selon une ligne i quelconque, par la formule

$$\det[a_{ij}] = (-1)^{i+1}a_{i1}\Delta_{i1} + (-1)^{i+2}a_{i2}\Delta_{i2} + \cdots + (-1)^{i+n}a_{in}\Delta_{in}\,.$$

Puisque transposer la matrice ne change pas le déterminant, on peut aussi développer selon une colonne j, avec la formule

$$\det[a_{ij}] = (-1)^{1+j} a_{1j}\Delta_{1j} + (-1)^{2+j} a_{2j}\Delta_{2j} + \cdots + (-1)^{n+j} a_{nj}\Delta_{nj}\,.$$

Les signes qu'il faut affecter aux coefficients le long de la ligne ou de la colonne choisie pour développer sont donnés par le tableau :

$$\begin{matrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{matrix}$$

Le premier signe en haut à gauche est + et l'on garnit le tableau en changeant de signe chaque fois qu'on passe à la colonne ou à la ligne suivante.

2.4 Calcul d'un déterminant

Le déterminant garde la même valeur si l'on ajoute à un vecteur-colonne, ou à un vecteur-ligne, une combinaison linéaire des autres. Par cette opération, on peut faire apparaître des zéros sur une même ligne ou sur une même colonne, comme lorsqu'on échelonne des vecteurs. Quand on a obtenu une ligne ou une colonne contenant peu de coefficients non nuls, on peut raisonnablement développer le déterminant selon cette ligne ou cette colonne.

Exemple. Calculons le déterminant $d = \begin{vmatrix} 2-x & 1 & 2 & -1 \\ 1 & \frac{1}{2}-x & 1 & -\frac{1}{2} \\ 2 & 1 & 2-x & -1 \\ -1 & -\frac{1}{2} & -1 & \frac{1}{2}-x \end{vmatrix}$.

Soustrayons la troisième colonne à la première, puis ajoutons la quatrième à la seconde :

$$d = \begin{vmatrix} -x & 1 & 2 & -1 \\ 0 & \frac{1}{2}-x & 1 & -\frac{1}{2} \\ x & 1 & 2-x & -1 \\ 0 & -\frac{1}{2} & -1 & \frac{1}{2}-x \end{vmatrix} = \begin{vmatrix} -x & 0 & 2 & -1 \\ 0 & -x & 1 & -\frac{1}{2} \\ x & 0 & 2-x & -1 \\ 0 & -x & -1 & \frac{1}{2}-x \end{vmatrix}$$

Ajoutons la première ligne à la troisième :

$$d = \begin{vmatrix} -x & 0 & 2 & -1 \\ 0 & -x & 1 & -\frac{1}{2} \\ 0 & 0 & 4-x & -2 \\ 0 & -x & -1 & \frac{1}{2}-x \end{vmatrix}$$

En développant selon la première colonne, il vient $d = -x\begin{vmatrix} -x & 1 & -\frac{1}{2} \\ 0 & 4-x & -2 \\ -x & -1 & \frac{1}{2}-x \end{vmatrix}$. Dans le déterminant de taille 3, soustrayons la première ligne de la troisième :

$$d = -x\begin{vmatrix} -x & 1 & -\frac{1}{2} \\ 0 & 4-x & -2 \\ 0 & -2 & 1-x \end{vmatrix}$$

Développons selon la première colonne :

$$d = (-x)(-x)\begin{vmatrix} 4-x & -2 \\ -2 & 1-x \end{vmatrix} = x^2\big[(4-x)(1-x) - (-2)(-2)\big] = x^2(x^2 - x) = x^3(x-1)\,.$$

Proposition

i) *Le déterminant d'une matrice triangulaire est le produit des coefficients diagonaux.*

ii) *Si A et B sont des matrices carrées de taille n et p, une matrice carrée de taille $n+p$ de la forme* $\begin{bmatrix} \boxed{A} & \boxed{C} \\ 0 & \boxed{B} \end{bmatrix}$ *a pour déterminant* $(\det A)(\det B)$.

Démonstration. Si $A = [a_{ij}]$ est une matrice triangulaire inférieure, le développement selon la première ligne est $\det M = a_{11}\Delta_{11}$, et Δ_{11} est un déterminant triangulaire inférieur ayant pour coefficients diagonaux $a_{22}, \ldots, a_{nn}$. On en déduit (i) en raisonnant par récurrence.

Soit $B \in \mathscr{M}_p(\mathbb{K})$ et $C \in \mathscr{M}_{n,p}(\mathbb{K})$. Pour toute matrice $M \in \mathscr{M}_n(\mathbb{K})$, posons $f(M) = \begin{vmatrix} \boxed{M} & \boxed{C} \\ 0 & \boxed{B} \end{vmatrix}$.

La fonction f possède les propriétés 1, 2 et 3 (par rapport aux colonnes de M). De plus, en développant le déterminant $\begin{vmatrix} \boxed{I_n} & \boxed{C} \\ 0 & \boxed{B} \end{vmatrix}$ selon la première colonne, il vient $f(I_n) = \det B$.

D'après la proposition page 148, on en déduit $f(M) = f(I_n)\det M = (\det B)(\det M)$ pour tout M. D'où le résultat en choisissant $M = A$. ■

2.5 Polynôme caractéristique d'une matrice carrée

Définition

Soit A une matrice carrée de taille n. Pour tout nombre $z \in \mathbb{K}$, on pose $C_A(z) = \det(A - zI_n)$. La fonction $z \mapsto C_A(z)$ est un polynôme de degré n, appelé *polynôme caractéristique de A*.

Exemple. Si $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, alors $\det(A - zI_2) = \begin{vmatrix} a-z & b \\ c & d-z \end{vmatrix} = (a-z)(d-z) - bc$ et il vient $C_A(z) = z^2 - (a+d)z + (ad - bc)$.

► Le polynôme caractéristique est de degré n et son terme constant est $C_A(0) = \det A$.

- On a $C_A = (-1)^n z^n + (-1)^{n-1}(\operatorname{tr} A)z^{n-1} + \cdots + \det A$, où $\operatorname{tr} A$ est la somme des coefficients diagonaux de A. Le nombre $\operatorname{tr} A$ s'appelle la *trace de A*.
- Les matrices A et ${}^tA$ ont même polynôme caractéristique : en effet, ${}^tA - zI_n = {}^t(A - zI_n)$ et deux matrices transposées ont même déterminant.

Voici une propriété importante du polynôme caractéristique. Rappelons que si $P = a_k z^k + \cdots + a_1 z + a_0$ est un polynôme, on a défini (page 134) la matrice $P(A)$ en posant $P(A) = a_k A^k + \cdots + a_1 A + a_0 I_n$.

Théorème de Cayley-Hamilton. *Pour toute matrice carrée A, la matrice $C_A(A)$ est nulle.*

Démonstration. Posons $A = [a_{ij}]$, donc $A - zI_n = \begin{bmatrix} a_{11}-z & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-z & \cdots & a_{2n} \\ \vdots & & \vdots & \\ a_{n1} & a_{n2} & \cdots & a_{nn}-z \end{bmatrix}$. Posons $D_j(z) = (-1)^{1+j}\Delta_{1j}$, où les Δ_{ij} sont les cofacteurs de $A - zI_n$. Alors $D_j(z)$ est un polynôme et par définition du polynôme caractéristique de A, on a

(1) $$C_A(z) = (a_{11} - z)D_1(z) + a_{12}D_2(z) + \cdots + a_{1k}D_k(z) + \cdots + a_{1n}D_n(z)$$

Remplaçons la première ligne par la k-ème, où $k \geqslant 2$: les cofacteurs Δ_{1j} sont inchangés, mais le déterminant obtenu est nul, car il a deux lignes égales :

(2) $$0 = a_{k1}D_1(z) + a_{k2}D_2(z) + \cdots + (a_{kk} - z)D_k(z) + \cdots + a_{kn}D_n(z)$$

En notant A_j la j-ème colonne de A, on a $A\boldsymbol{E_j} = A_j = a_{1j}\boldsymbol{E_1} + a_{2j}\boldsymbol{E_2} + \cdots + a_{jj}\boldsymbol{E_j} + \cdots + a_{nj}\boldsymbol{E_n}$ ou encore

(3) $$0 = a_{1j}\boldsymbol{E_1} + a_{2j}\boldsymbol{E_2} + \cdots + (a_{jj}I_n - A)\boldsymbol{E_j} + \cdots + a_{nj}\boldsymbol{E_n}$$

Écrivons ces égalités pour $j = 1, 2, \ldots, n$ puis multiplions la première par la matrice $D_1(A)$, la deuxième par $D_2(A)$, la j-ème par $D_j(A)$ et la n-ième par $D_n(A)$:

$$\begin{array}{llccccccc}
D_1(A)\times & 0 = & (a_{11}I_n - A)\boldsymbol{E_1} & + & a_{21}\boldsymbol{E_2} & + & \cdots & + & a_{n1}\boldsymbol{E_n} \\
D_2(A)\times & 0 = & a_{12}\boldsymbol{E_1} & + & (a_{22}I_n - A)\boldsymbol{E_2} & + & \cdots & + & a_{n2}\boldsymbol{E_n} \\
\cdots & \cdots & & & & & \cdots & & \\
D_n(A)\times & 0 = & a_{1n}\boldsymbol{E_1} & + & a_{2n}\boldsymbol{E_2} & + & \cdots & + & (a_{nn}I_n - A)\boldsymbol{E_n}
\end{array}$$

En ajoutant ces égalités en colonne, on obtient l'égalité matricielle

(4) $$0 = M_1\boldsymbol{E_1} + M_2\boldsymbol{E_2} + \cdots + M_n\boldsymbol{E_n}$$

où $M_k = D_1(A)a_{k1} + D_2(A)a_{k2} + \cdots + D_k(A)(a_{kk}I_n - A) + \cdots + D_n(A)a_{kn}$.
Posons $P_k(z) = a_{k1}D_1(z) + a_{k2}D_2(z) + \cdots + (a_{kk} - z)D_k(z) + \cdots + a_{kn}D_n(z)$.
Le polynôme P_k vérifie $P_k(A) = M_k$, car A et $D_k(A)$ commutent. D'après (1), on a $P_1 = C_A$ et d'après (2), il vient $P_k = 0$ pour tout $k = 2, \ldots, n$. L'égalité (4) s'écrit donc

$$0 = P_1(A)\boldsymbol{E_1} + P_2(A)\boldsymbol{E_2} + \cdots + P_n(A)\boldsymbol{E_n} = C_A(A)\boldsymbol{E_1}\,.$$

Cela veut dire que la première colonne de la matrice $C_A(A)$ est nulle. Pour montrer que la i-ième colonne de $C_A(A)$ est nulle, on remplace, dans la matrice $A - zI_n$, la i-ème ligne par la k-ième, où $k \neq i$, et l'on suit le même raisonnement. ■

Exemple. Pour une matrice de taille 2, le théorème de Cayley-Hamilton affirme que l'on a $A^2 - (\operatorname{tr} A)A + (\det A)I_2 = 0$. Cette identité se vérifie directement.

2.6 Applications des déterminants

Caractérisation des bases de $\mathbb{K}^p$

Des vecteurs $u_1, u_2, \ldots, u_p$ de $\mathbb{K}^p$ forment une base de $\mathbb{K}^p$ si et seulement s'ils sont indépendants (propriété 2, page 120), c'est-à-dire si et seulement si la matrice de colonnes $u_1, \ldots, u_p$ est inversible. On a donc l'équivalence suivante.

Des vecteurs $u_1,u_2,\ldots,u_p$ *forment est une base de* $\mathbb{K}^p$ *si et seulement si* $\det(u_1,u_2,\ldots,u_p)\neq 0$.

Équation d'un hyperplan

Soit H l'hyperplan de $\mathbb{K}^p$ engendré par les $p-1$ vecteurs indépendants $u_1,u_2,\ldots,u_{p-1}$. Cela veut dire qu'un vecteur X appartient à H si et seulement s'il est combinaison linéaire des u_i. D'après le théorème page 105, cette condition signifie que les vecteurs $u_1, u_2, \ldots, u_{p-1}, X$ ne sont pas indépendants. On a donc l'équivalence :

$$X \in H \iff \det(u_1, u_2, \ldots, u_{p-1}, X) = 0$$

Il suffit de poser $X = (x_1, \ldots, x_p)$ et de développer le déterminant pour avoir l'équation de H sous la forme $a_1x_1 + \cdots + a_px_p = 0$.

Équation d'un plan affine de $\mathbb{R}^3$. Soient A, B, C trois points de $\mathbb{R}^3$ non alignés et soit P le plan affine passant par A, B et C. Un point M appartient à P si et seulement si le vecteur $\overrightarrow{CM}$ est combinaison linéaire des vecteurs indépendants $\overrightarrow{AB}$ et $\overrightarrow{BC}$. En notant (a_1, a_2, a_3) les coordonnées de A et (b_1, b_2, b_3) celles de B, une équation de P est donc

$$\begin{vmatrix} b_1-a_1 & c_1-b_1 & x-c_1 \\ b_2-a_2 & c_2-b_2 & y-c_2 \\ b_3-a_3 & c_3-b_3 & z-c_3 \end{vmatrix} = 0$$

Estimation du rang d'une matrice

Soit A une matrice à p lignes et n colonnes. Supposons que r est un entier inférieur ou égal à p et n et que la matrice U formée des r premières lignes et des r premières colonnes de A a son déterminant non nul :

$$A = \begin{bmatrix} \boxed{U} & * \\ * & * \end{bmatrix} \quad \text{et } \det U \neq 0.$$

Par hypothèse, il n'y a aucune relation linéaire entre les colonnes de U, donc *a fortiori* il n'y en a aucune entre les r premières colonnes de A. Les r premières colonnes de A sont donc indépendantes et par suite, on a $\operatorname{rg} A \geqslant r$.

Plus généralement, si en sélectionnant r lignes et r colonnes de A, on peut former une matrice de déterminant non nul, alors le rang de A est supérieur ou égal à r.

Élimination polynomiale

On a parfois besoin de savoir si deux polynômes ont une racine complexe commune. Nous allons voir que les déterminants permettent d'exprimer cette propriété par une relation entre les coefficients des polynômes : cela est utile, car on ne sait pas, en général, calculer les racines des polynômes.

Soient A et B des polynômes non constants à coefficients réels ou complexes.

► Supposons que A et B ont une racine commune $a \in \mathbb{C}$. On a donc $A = (z-a)P$ et $B=(z-a)Q$, où P et Q sont des polynômes. Il vient $QA-PB=Q(z-a)P-P(z-a)Q=0$, donc $QA = PB$, avec $\deg P = (\deg A) - 1$ et $\deg Q = (\deg B) - 1$.

► Supposons que A et B n'ont pas de racine commune, c'est-à-dire sont étrangers. D'après la proposition page 50, il existe des polynômes U et V tels que $AU + BV = 1$. Supposons que P et Q sont des polynômes non nuls tels que $QA = PB$. On a $P = P(AU+BV) = PAU+PBV = PAU+QAV = A(PU+QV)$, donc $\deg P \geqslant \deg A$. Puisque $\deg A + \deg Q = \deg B + \deg P$, il s'ensuit $\deg Q \geqslant \deg B$.

Ce raisonnement montre que les polynômes A et B ont une racine commune dans $\mathbb{C}$ si et seulement s'il existe des polynômes non nuls P et Q tels que $QA = PB$, $\deg P < \deg A$ et $\deg Q < \deg B$.

Posons $A = a_p z^p + a_{p-1}z^{p-1} + \cdots + a_1 z + a_0$ et $B = b_q z^q + b_{q-1}z^{q-1} + \cdots + b_1 z + b_0$, où les coefficients a_p et b_q sont différents de 0.

Un polynôme Q de degré strictement inférieur à q est une somme de monômes $\beta_i z^i$ où $0 \leqslant i \leqslant q-1$, donc $QA = \sum_{i=0}^{q-1} \beta_i z^i A$. De même, si $\deg P < p$, alors $PB = \sum_{j=0}^{p-1} \gamma_j z^j B$. Les polynômes QA et PB sont de degré inférieur ou égal à $p+q-1$.

À tout polynôme $T = t_0 + t_1 z + \cdots + t_{p+q-1}z^{p+q-1}$, associons le vecteur des coefficients $v(T) = (t_0, t_1, \ldots, t_{p+q-1})$. On a donc $v(T) \in \mathbb{C}^{p+q}$.

Écrivons en ligne les vecteurs associés aux polynômes $A, zA, \ldots, z^{q-1}A$:

$$\begin{array}{rcccccccccc}
v(A) & : & a_0 & a_1 & a_2 & \cdots & a_p & 0 & 0 & \cdots & 0 \\
v(zA) & : & 0 & a_0 & a_1 & \cdots & a_{p-1} & a_p & 0 & \cdots & 0 \\
\cdots & & & & & \cdots & & & & & \\
v(z^{q-1}A) & : & 0 & \cdots & 0 & a_0 & a_1 & a_2 & \cdots & a_{p-1} & a_p
\end{array}$$

Dans cette disposition, les coefficients de A sont décalés d'un rang à chaque ligne ; il y a q lignes et $p+q$ colonnes.

En faisant de même avec les p polynômes $B, zB, \ldots, z^{p-1}B$ et en superposant les deux tableaux, on obtient la matrice carrée de taille $p+q$

$$\begin{bmatrix}
a_0 & a_1 & a_2 & \cdots & a_p & 0 & \cdots & & 0 \\
0 & a_0 & a_1 & \cdots & a_{p-1} & a_p & 0 & \cdots & 0 \\
0 & 0 & a_0 & a_1 & \cdots & a_{p-1} & a_p & 0 \cdots & 0 \\
\vdots & & & & & & \vdots & & \\
0 & \cdots & & & 0 & a_0 & a_1 & \cdots a_{p-1} & a_p \\
b_0 & b_1 & b_2 & \cdots & \cdots & b_q & 0 & \cdots & 0 \\
0 & b_0 & b_1 & \cdots & \cdots & b_{q-1} & b_q & 0 \cdots & 0 \\
\vdots & & & & & & & \vdots & \\
0 & \cdots & & 0 & b_0 & b_1 & \cdots & \cdots b_{q-1} & b_q
\end{bmatrix}$$

Le déterminant $R(A,B)$ de cette matrice s'appelle le *résultant* des polynômes A et B.

S'il existe des polynômes non nuls P et Q tels que $QA = PB$, $\deg P < \deg A$ et $\deg Q < \deg B$, alors avec les notations précédentes, on obtient

$$v(QA) = \sum_{i=0}^{q-1} \beta_i v(z^i A) = v(PB) = \sum_{j=0}^{p-1} \gamma_j v(z^j B)$$

L'égalité $v(QA) - v(PB) = 0$ est une relation linéaire entre les vecteurs-lignes de la matrice, donc le résultant est nul. Réciproquement, s'il existe une combinaison linéaire $\sum_{i=0}^{q-1} \beta_i v(z^i A) - \sum_{j=0}^{p-1} \gamma_j v(z^j B) = 0$ à coefficients non tous nuls, alors les polynômes $Q = \sum_{i=0}^{q-1} \beta_i z^i$ et $P = \sum_{j=0}^{p-1} \gamma_j z^j$ sont non nuls et vérifient $v(QA) = v(PB)$, donc $QA = PB$. On a ainsi la proposition suivante.

Proposition. *Des polynômes non constants A et B ont une racine complexe commune si et seulement si leur résultant $R(A,B)$ est nul.*

Si l'on peut éliminer la variable z entre les deux équations $A(z) = 0$ et $B(z) = 0$, on obtient une condition équivalente à $R(A,B) = 0$.

Exemple. Considérons un circuit électrique composé d'une résistance ρ, d'une inductance L et d'une capacité C montées en série. L'intensité en régime libre est une fonction $\iota = \dfrac{dq}{dt}$ du temps, où la charge $q(t)$ de la capacité satisfait l'équation différentielle $L\dfrac{d^2q}{dt^2} + \rho\dfrac{dq}{dt} + \dfrac{q}{C} = 0$ (voir au chapitre 15).
Si les racines du polynôme $A = Lz^2 + \rho z + 1/C$ sont les nombres complexes $r \pm \omega \boldsymbol{i}$, où $\omega \neq 0$, les solutions sont $q(t) = ae^{rt}\cos(\omega t + \phi)$ et l'intensité présente des oscillations amorties (car $r = -\rho/2L$ est négatif). Si le polynôme A possède deux racines réelles λ et μ, les solutions sont $q(t) = ae^{\lambda t} + be^{\mu t}$: le régime est apériodique et l'intensité tend encore vers 0 quand t tend vers l'infini, car λ et μ sont négatifs (page 52).
La condition pour que deux circuits oscillants aient un régime commun est $R(A,B)=0$, où A et B sont les polynômes associés aux équations différentielles. En posant

$B = L'z^2 + \rho' z + 1/C'$, il vient

$$R(A,B) = \begin{vmatrix} 1/C & \rho & L & 0 \\ 0 & 1/C & \rho & L \\ 1/C' & \rho' & L' & 0 \\ 0 & 1/C' & \rho' & L' \end{vmatrix}$$

et la condition est $\left(\frac{1}{L'C'} - \frac{1}{LC}\right)^2 + \frac{1}{LL'}\left(\frac{\rho'}{C} - \frac{\rho}{C'}\right)\left(\frac{\rho}{L} - \frac{\rho'}{L'}\right) = 0$.

Exercices

1. Montrer que si $r \neq \pm 1$, l'inverse de la matrice $A = \begin{bmatrix} 1 & r & r^2 & r^3 \\ r & 1 & r & r^2 \\ r^2 & r & 1 & r \\ r^3 & r^2 & r & 1 \end{bmatrix}$ est la matrice $A^{-1} = \frac{1}{r^2-1}\begin{bmatrix} -1 & r & 0 & 0 \\ r & -r^2-1 & r & 0 \\ 0 & r & -r^2-1 & r \\ 0 & 0 & r & -1 \end{bmatrix}$. En déduire l'inverse de $B = \begin{bmatrix} 1 & -2 & 0 & 0 \\ -2 & 5 & -2 & 0 \\ 0 & -2 & 5 & -2 \\ 0 & 0 & -2 & 1 \end{bmatrix}$.

@ **2. a)** Calculer le déterminant de la matrice $A = \begin{bmatrix} b & 1 & a \\ 1 & a & 1 \\ a & 1 & b \end{bmatrix}$. Montrer que si $a = b$, ce déterminant est nul. Mettre en facteur $a - b$ dans l'expression de $\det A$.

b) Calculer le rang de la matrice A en fonction des valeurs de a et de b.

3. Soit L une matrice-ligne à n colonnes et C une matrice-colonne à n lignes. Montrer que si L ou C n'est pas nulle, la matrice produit CL (qui est carrée de taille n) est de rang 1.

4. On considère un système linéaire $AX = K$, où A est une matrice et K une matrice-colonne non nulle. Supposons que X_1 et X_2 sont des solutions. Comment faut-il choisir les nombres a_1 et a_2 pour que $a_1X_1 + a_2X_2$ soit solution ?

5. Matrices de Pauli. En Mécanique quantique, on utilise les matrices de Pauli. Ce sont les matrices à coefficients complexes définies par

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad , \quad \sigma_y = \begin{bmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{bmatrix} \quad , \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

a) Montrer que ces matrices ont pour déterminant -1 et pour trace 0.

b) Montrer que l'on a $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = I_2$, $\sigma_x\sigma_y = -\sigma_y\sigma_x = \boldsymbol{i}\sigma_z$ et $\sigma_x\sigma_y - \sigma_y\sigma_x = 2\boldsymbol{i}\sigma_z$.

c) Démontrer l'égalité $\sigma_x\sigma_y\sigma_z = \boldsymbol{i}\, I_2$.

d) Soit $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Trouver des nombres p, q_x, q_y, q_z tels que $M = pI_2 + q_x\sigma_x + q_y\sigma_y + q_z\sigma_z$. Vérifier que l'on a $p = (1/2)\,\mathrm{tr}(M)$, $q_x = (1/2)\,\mathrm{tr}(M\sigma_x)$, $q_y = (1/2)\,\mathrm{tr}(M\sigma_y)$ et $q_z = (1/2)\,\mathrm{tr}(M\sigma_z)$.

e) Les matrices de Pauli ont toutes le même polynôme caractéristique : calculer ce polynôme.

@ **6. Tableaux d'achats-ventes.** On considère une économie en système fermé, formée de n secteurs d'activité $S_1, \ldots, S_n$. Notons x_i le volume des ventes du secteur S_i et $x_{i,j}$ la part de ces ventes utilisée en achat par le secteur S_j, où $j \neq i$. L'équilibre du secteur S_1 s'écrit donc $x_{1,2} + x_{1,3} + \cdots + x_{1,n} = x_1$.

a) Faisons l'hypothèse (raisonnable) que le montant des biens fournis par le secteur S_i au secteur S_j est proportionnel à la production-vente de S_j : cela se traduit par l'existence d'un coefficient constant $a_{i,j}$ tel que $x_{i,j} = a_{i,j}x_j$. Posons pour simplifier $a_{i,i} = 0$ et notons A la matrice carrée de taille n : $A = [a_{i,j}]$. Posons $M = I_n - A$ et $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Montrer que l'on a $MX = 0$.

b) Notons $p_1, \ldots, p_n$ les prix des produits des différents secteurs. En supposant qu'il n'y a pas de profit, le prix de vente d'une unité du secteur S_1 est égal à la somme des prix d'achats auprès des autres secteurs, ce qui s'écrit $p_1 = p_2 a_{2,1} + \cdots + p_n a_{n,1}$. Notons $P = [\, p_1 \;\; p_2 \;\; \cdots \;\; p_n \,]$ la matrice-ligne des prix. Montrer que l'on a $PM = 0$.

c) On suppose $n=3$, $a_{1,2}=1/5$, $a_{1,3}=2/3$, $a_{2,1}=3/5$, $a_{2,3}=8/15$, $a_{3,1}=1/3$, $a_{3,2}=2/3$.

(i) Calculer le déterminant et le rang de M.

(ii) Résoudre l'équation linéaire $MX = 0$ et calculer les valeurs relatives x_2/x_1 et x_3/x_1.

(iii) Résoudre l'équation linéaire $PM = 0$ (où l'inconnue est le vecteur-ligne P) et calculer les valeurs relatives p_2/p_1 et p_3/p_1.

@ **7. Équilibre de la consommation.** Considérons une économie ayant trois secteurs d'activité S_1, S_2, S_3 produisant des quantités de biens x_1, x_2, x_3. Comme dans l'exercice précédent, la fraction de x_i utilisée par le secteur S_j est $a_{i,j}x_j$, où (pour $i \neq j$), $a_{i,j}$ est un coefficient caractéristique de cette économie. Notons y_1 la demande des consommateurs pour le produit du secteur S_1, y_2 la demande concernant S_2 et y_3 la demande concernant S_3.

a) Écrire les équations qui expriment l'équilibre entre production et consommation dans chacun des secteurs.

b) On suppose $a_{1,2} = 1/4$, $a_{1,3} = 1/2$, $a_{2,1} = 1/5$, $a_{2,3} = 1/2$, $a_{3,1} = 2/5$ et $a_{3,2} = 3/4$. Écrire ces équations sous la forme $MX = Y$, où $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ et $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$.

c) Montrer que la matrice M est inversible et calculer x_1, x_2, x_3 au moyen de y_1, y_2, y_3.

d) On suppose que la demande des consommateurs pour le produit du secteur S_1 augmente d'une unité. Pour répondre à cette demande, de combien faudra-t-il augmenter les productions dans chaque secteur ?

8. Un déterminant de matrice antisymétrique. Montrer la formule

$$\begin{vmatrix} 0 & x & y & z \\ -x & 0 & t & u \\ -y & -t & 0 & v \\ -z & -u & -v & 0 \end{vmatrix} = (xv - yu + zt)^2$$

9. Soit le polynôme $P = z^3 + z^2 + az + 9$, où a est un nombre réel ou complexe.

a) Montrer que le résultant des polynômes P et P' est $(4a + 39)(a^2 - 10a + 57)$.

b) Pour quelles valeurs de a le polynôme P a-t-il une racine multiple ?

c) Trouver toutes les racines de P dans le cas $a = -39/4$.

@10. Calculer les quatre valeurs du nombre complexe q pour lesquelles le système d'équations $\begin{cases} 2z^3 + q & = 0 \\ z^2 + qz + 1 & = 0 \end{cases}$ a une solution au moins. Résoudre le système pour ces valeurs de q.

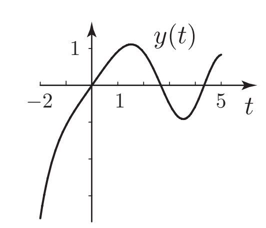
11. Approche numérique d'une équation différentielle. Considérons l'équation différentielle d'Airy $(*)$ $\;y''(t) + ty(t) = 0$ (utilisée en Physique) et tâchons d'approximer les solutions $y(t)$. Faisons les calculs sur l'intervalle $[-2, 5]$ en des points t_i régulièrement espacés d'une quantité $h > 0$.

a) Montrer que l'on peut approcher les valeurs $y(t_i)$ par des nombres y_i vérifiant les équations $y_{i-1} - 2y_i + y_{i+1} + h^2 t_i y_i = 0$.

b) Intéressons-nous à la seule solution $y(t)$ de $(*)$ qui passe par l'origine avec une dérivée a, donc $y(0) = 0$ et $y'(0) = a$. Choisissons la subdivision de manière que 0 est l'un des points t_i : il y a un indice p tel que $t_p = y_p = 0$. Montrer que dans notre approximation, la condition $y'(0) = a$ peut s'exprimer par $y_{p+1} - y_{p-1} = 2ah$.

c) Écrire les huit équations obtenues en choisissant $h=1$ et les points $t_i=i$ pour $-2 \leqslant i \leqslant 5$. En déduire que les valeurs approchées y_i de $y(t_i)$ vérifient $y_0=0$, $y_1=2a+y_{-1}$ et

$$\begin{bmatrix} 1 & -3 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_{-2} \\ y_{-1} \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ -2a \\ 2a \\ -2a \\ 0 \\ 0 \end{bmatrix}$$



d) Résoudre les équations précédentes. Dans le cas $a = 1$, dessiner la ligne brisée passant par les points (t_i, y_i).
Ci-contre le graphe de la solution $t \mapsto y(t)$ sur l'intervalle $[-2, 5]$, pour $a = 1$.

@12. Posons $q(x, y) = ax^2 + bxy + cy^2$, $A = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}$ et $P = \begin{bmatrix} p & -uc \\ ua & p+ub \end{bmatrix}$, où a, b, c, p, u sont des nombres réels.

a) Montrer que l'on a $({}^tP)AP = (p^2 + bup + acu^2)A$.

b) Montrer que pour tout vecteur-colonne $X = \begin{bmatrix} x \\ y \end{bmatrix}$, on a $({}^tX)AX = ax^2 + bxy + cy^2$.

c) On considère la transformation linéaire définie par $X' = PX$. En utilisant la question précédente, montrer que si l'on pose $X = \begin{bmatrix} x \\ y \end{bmatrix}$ et $X' = \begin{bmatrix} x' \\ y' \end{bmatrix}$, alors on a $q(x', y') = (p^2 + bup + acu^2)q(x, y)$.

d) On suppose désormais $a = 3$ et $b = c = -2$. Calculer le nombre $p^2 + bup + acu^2$ pour $u = 3$ et $p = 11$. En déduire l'identité $q(x', y') = q(x, y)$.

e) Montrer que les itérés d'un vecteur $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$ par la transformation $X \mapsto PX$ sont les vecteurs $\begin{bmatrix} x_n \\ y_n \end{bmatrix}$ définis par les formules $x_{n+1} = 11x_n + 6y_n$, $y_{n+1} = 9x_n + 5y_n$. Montrer que si l'on pose $(x_0, y_0) = (2, 1)$ ou bien $(x_0, y_0) = (4, 3)$, alors x_n et y_n sont des entiers solutions de l'équation $3x^2 - 2xy - 2y^2 = 6$.

f) Étudions la courbe (C) d'équation $q(x,y)=6$ au moyen d'un changement de repère.

(i) Calculer les racines t_1 et t_2 de l'équation $q(t, 1) = 0$ et montrer que l'on a $q(x, y) = 3(x - t_1y)(x - t_2y)$.

(ii) Montrer que les vecteurs $U_1 = (t_1, 1)$ et $U_2 = (t_2, 1)$ forment une base de $\mathbb{R}^2$ et que si l'on note (X, Y) les coordonnées d'un point de $\mathbb{R}^2$ dans cette base, la courbe (C) a pour équation $XY = -9/14$. En déduire que (C) est une hyperbole dont les asymptotes sont portées par les vecteurs U_1, U_2. Dessiner la courbe (C).

D'après la question (d), la transformation linéaire $X \mapsto PX$ envoie les points de (C) dans des points de (C) : les itérés d'un point de (C) sont tous sur (C).

@13. L'inverse d'une matrice est un polynôme en la matrice. Posons $M = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & 2 \\ -1 & 2 & 1 \end{bmatrix}$.

a) Calculer le polynôme caractéristique de M. Déduire du théorème de Cayley-Hamilton que l'on a $M(M^2 - 9I_3) + 9I_3 = 0$ et $M^{-1} = I_3 - (1/9)M^2$.

b) Montrer que pour tout nombre z, on a $M^3 - zI_3 = 9\big[M - (1+z/9)I_3\big]$. En déduire que le polynôme caractéristique de M^3 est $9^3C_M(1+z/9)$, où C_M est le polynôme caractéristique de M.

Chapitre 6

Espaces vectoriels et applications linéaires

1. Espaces vectoriels

Au chapitre 4, nous avons utilisé les combinaisons linéaires de vecteurs de $\mathbb{K}^n$. Voici d'autres exemples où l'on définit de manière naturelle la somme et le produit par un nombre.

Exemples

a) Soit I une partie de $\mathbb{R}$. Si f et g sont des fonctions de I dans $\mathbb{R}$,

- la somme $f+g : I \to \mathbb{R}$ est la fonction $t \mapsto f(t)+g(t)$;
- pour tout nombre réel λ, le produit $\lambda f : I \to \mathbb{R}$ est la fonction $t \mapsto \lambda f(t)$.

b) Si E est un ensemble quelconque et si f et g sont des fonctions de E dans $\mathbb{K}$, ou bien de E dans $\mathbb{K}^p$, on définit de même la somme $f+g$ et le produit λf par un nombre appartenant à $\mathbb{K}$.
Ainsi, la somme de deux suites (u_n) et (v_n) est la suite de terme général $u_n + v_n$ et pour λ réel (ou complexe), le produit par λ de la suite (u_n) est la suite de terme général λu_n.

c) Si A et B sont des matrices à p lignes et n colonnes, leur somme $A+B$ et le produit λA par un scalaire sont des matrices à p lignes et n colonnes.

d) La somme de deux polynômes à coefficients dans $\mathbb{K}$ est un polynôme à coefficients dans $\mathbb{K}$, de même que le produit d'un polynôme par un scalaire de $\mathbb{K}$. Dans l'ensemble $\mathbb{P}(\mathbb{K})$ des polynômes à coefficients dans $\mathbb{K}$, on dispose donc d'une addition et du produit par un scalaire.
Soit d un entier positif et soit $\mathbb{P}_d(\mathbb{K})$ l'ensemble des polynômes de la forme $a_d z^d + a_{d-1} z^{d-1} + \cdots + a_1 z + a_0$: ce sont les polynômes de degré au plus d et le polynôme nul. Le degré d'une somme de polynômes est inférieur ou égal au maximum des degrés, et le degré d'un polynôme ne change pas quand on le

multiplie par un scalaire non nul ; dans l'ensemble $\mathbb{P}_d(\mathbb{K})$, on a donc encore les opérations d'addition et de produit par un scalaire.

Dans ces exemples, les règles de calcul concernant la somme et le produit par un scalaire sont celles qu'on pratique dans l'espace vectoriel $\mathbb{K}^p$. Énonçons-les dans le cas général d'un ensemble V d'éléments appelés *vecteurs*.

i) Il existe un vecteur nul $\mathbf{0}$ tel que $u + \mathbf{0} = \mathbf{0} + u = u$ pour tout vecteur $u \in V$.

ii) On a $u + v = v + u$ et $u + (v + w) = (u + v) + w$ pour tous vecteurs $u, v, w \in V$.

iii) Pour tout vecteur $u \in V$, il existe un vecteur opposé, noté $-u$, tel que $u + (-u) = (-u) + u = \mathbf{0}$.

iv) Pour tous vecteurs $u, v \in V$ et pour tous scalaires $\lambda, \mu \in \mathbb{K}$, on a

$$\lambda(u + v) = \lambda u + \lambda v \ , \quad (\lambda + \mu)u = \lambda u + \mu u \ , \quad \lambda(\mu u) = (\lambda\mu)u \ , \quad 1\, u = u \, .$$

Un ensemble V muni d'opérations ayant ces propriétés s'appelle un $\mathbb{K}$-*espace vectoriel*. On démontre facilement que dans un $\mathbb{K}$-espace vectoriel V, on a aussi les relations

$$(-1)u = -u \ , \quad 0\, u = \mathbf{0} \ , \quad \lambda\, \mathbf{0} = \mathbf{0}$$

et l'équivalence $\lambda u = \mathbf{0} \iff (\lambda = 0 \ \text{ ou } \ u = \mathbf{0})$.

Dans un $\mathbb{K}$-espace vectoriel, on définit les combinaisons linéaires de vecteurs et les notions de vecteurs indépendants, de vecteurs qui engendrent et de base comme nous l'avons fait dans $\mathbb{K}^p$.

Définitions

Soit V un $\mathbb{K}$-espace vectoriel et soient $u_1, \ldots, u_n$ des vecteurs appartenant à V.

- Un vecteur $\lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_n u_n$ s'appelle une *combinaison linéaire* des vecteurs $u_1, \ldots, u_n$.
- Les vecteurs $u_1, \ldots, u_n$ sont *indépendants* si l'équation linéaire $x_1 u_1 + x_2 u_2 + \cdots + x_n u_n = \mathbf{0}$ n'a que la solution $x_1 = x_2 = \cdots = x_n = 0$.
- Si tout vecteur de V est combinaison linéaire des vecteurs $u_1 \ldots, u_n$, on dit que $u_1, \ldots, u_n$ *engendrent* V.
- Les vecteurs $u_1, \ldots, u_n$ forment une *base* de V si, pour tout vecteur $v \in V$, l'équation linéaire $x_1 u_1 + x_2 u_2 + \cdots + x_n u_n = v$ possède une unique solution.

Des vecteurs $u_1, \ldots, u_n$ de l'espace vectoriel V forment une base de V si et seulement s'ils engendrent V et sont indépendants.

Définition

Soit V un $\mathbb{K}$-espace vectoriel et soit W une partie de V. On dit que W est un *sous-espace vectoriel* de V si le vecteur nul appartient à W et si toute combinaison linéaire d'éléments de W appartient à W.

Si W est un sous-espace vectoriel de V, alors W, muni des mêmes opérations que V, est un $\mathbb{K}$-espace vectoriel. Voici des exemples d'espaces vectoriels.

Des espaces vectoriels de fonctions

a) Étant donné un ensemble E, l'ensemble de toutes les applications de E dans $\mathbb{K}$ est un $\mathbb{K}$-espace vectoriel, le vecteur nul étant la fonction $\mathbf{0} : E \to \mathbb{K}$ définie par $\mathbf{0}(x) = 0$ quel que soit $x \in E$.

b) Prenons $E = [a, b]$ (avec $a < b$) et posons $f_n(x) = x^n$ pour tout entier $n \geqslant 0$ (pour $n = 0$, on a par convention $f_0(x) = x^0 = 1$). La combinaison linéaire $g = \lambda_n f_n + \lambda_{n-1} f_{n-1} + \cdots + \lambda_1 f_1 + \lambda_0 f_0$ est la fonction définie par
$$g(x) = \lambda_n x^n + \lambda_{n-1} x^{n-1} + \cdots + \lambda_1 x + \lambda_0 \text{ , pour tout } x \in [a, b].$$
Puisque g est une fonction polynôme, elle n'est nulle que si tous ses coefficients sont nuls. Dans l'espace vectoriel des applications de $[a, b]$ dans $\mathbb{R}$, les vecteurs $f_0, f_1, \ldots, f_n$ sont donc indépendants. Comme cela est vrai quel que soit n, on peut trouver, dans cet espace, des familles de vecteurs indépendants comportant autant d'éléments que l'on veut.

c) L'ensemble des suites (u_n) à termes réels est un $\mathbb{R}$-espace vectoriel. Une combinaison linéaire de suites qui tendent vers 0 tend vers 0 : l'ensemble des suites réelles qui tendent vers 0 est donc un sous-espace vectoriel.

d) Si $f, g : \mathbb{R} \to \mathbb{R}$ sont des fonctions dérivables, leur somme $t \mapsto f(t) + g(t)$ est dérivable ainsi que la fonction $t \mapsto \lambda f(t)$: les fonctions dérivables forment donc un sous-espace vectoriel de l'espace vectoriel des fonctions de $\mathbb{R}$ dans $\mathbb{R}$.

Des espaces vectoriels de polynômes

- L'ensemble $\mathbb{P}(\mathbb{K})$ des polynômes à coefficients dans $\mathbb{K}$ est un $\mathbb{K}$-espace vectoriel. En faisant des combinaisons linéaires d'un nombre fini de polynômes, on n'obtiendra pas de polynômes de degré supérieur au plus grand degré des polynômes employés : l'espace vectoriel des polynômes ne peut donc pas être engendré par un nombre fini de vecteurs.
- Donnons-nous un entier d positif ou nul. Les polynômes de degré au plus d sont de la forme $a_0 + a_1 z + \cdots + a_d z^d$: avec le polynôme nul, ils constituent un sous-espace vectoriel de $\mathbb{P}(\mathbb{K})$, noté $\mathbb{P}_d(\mathbb{K})$. Les $d+1$ polynômes $1, z, \ldots, z^d$ forment une base de $\mathbb{P}_d(\mathbb{K})$.

Des espaces vectoriels de matrices

a) L'ensemble $\mathscr{M}_{p,n}(\mathbb{K})$ des matrices à p lignes, n colonnes et à coefficients dans $\mathbb{K}$ est un espace vectoriel. Le vecteur nul est la matrice dont tous les coefficients sont nuls. Considérons les matrices $E_{i,j}$ dont tous les coefficients sont nuls, sauf celui en position i, j qui vaut 1. Par exemple, pour $p = n = 2$, on a $E_{1,2} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $E_{2,2} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ et $xE_{1,1} + yE_{1,2} + zE_{2,1} + tE_{2,2} = \begin{bmatrix} x & y \\ z & t \end{bmatrix}$.

Plus généralement, les np matrices $E_{i,j}$ forment une base (dite canonique) de l'espace vectoriel $\mathscr{M}_{p,n}(\mathbb{K})$, car pour toute matrice $A=[a_{ij}]$, on a $A=\sum_{i,j} a_{ij}E_{i,j}$.

b) La somme de deux matrices carrées symétriques est symétrique et si l'on multiplie une matrice symétrique par un scalaire, elle reste symétrique. L'ensemble des matrices symétriques de taille n est donc un sous-espace vectoriel de $\mathscr{M}_n(\mathbb{K})$.
Pour $n=2$, la matrice symétrique générale s'écrit $\begin{bmatrix} x & y \\ y & t \end{bmatrix}=x\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}+y\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}+t\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$: les trois matrices $E_{1,1}, E_{1,2}+E_{2,1}, E_{2,2}$ forment donc une base de l'espace vectoriel des matrices symétriques de taille 2.
Plus généralement, l'espace vectoriel des matrices symétriques de taille n a pour base les matrices $E_{i,i}$ et $E_{j,k}+E_{k,j}$ pour $1 \leqslant i \leqslant n$ et $1 \leqslant j < k \leqslant n$; cette base comporte $n+\dfrac{n(n-1)}{2}=\dfrac{n(n+1)}{2}$ éléments.

1.1 Bases et dimension

Les résultats que nous avons démontrés dans l'espace vectoriel $\mathbb{K}^p$ restent vrais dans un $\mathbb{K}$-espace vectoriel possédant une base $e_1, \ldots, e_p$. Voici un théorème permettant de construire des bases.

Théorème de la base incomplète. *Soit V un $\mathbb{K}$-espace vectoriel engendré par un nombre fini de vecteurs $v_1, \ldots, v_q$ et soient $e_1, \ldots, e_n$ des vecteurs indépendants appartenant à V. Alors il existe une base de V de la forme $e_1, \ldots, e_n, f_1, \ldots, f_k$, où les f_i sont certains des vecteurs $v_1, \ldots, v_q$.*

Corollaire. *Tout espace vectoriel engendré par un nombre fini de vecteurs possède une base.*

Démonstration du théorème. Si chacun des vecteurs v_j est combinaison linéaire de e_1, $e_2, \ldots, e_n$, alors tout vecteur de V est combinaison de $e_1, e_2, \ldots, e_n$, car il est combinaison des v_j : dans ce cas, $e_1, e_2, \ldots, e_n$ est une base de V. Supposons que certains vecteurs v_j ne sont pas combinaison linéaire de $e_1, \ldots, e_n$ et complétons $e_1, e_2, \ldots, e_n$ avec le plus grand nombre possible de vecteurs $f_1, \ldots, f_k$ pris parmi les v_j et de telle manière que $e_1, \ldots, e_n, f_1, \ldots, f_k$ restent indépendants. Si v_j est un vecteur non sélectionné, alors par construction, $e_1, \ldots, e_n, f_1, \ldots, f_k, v_j$ ne sont pas indépendants, donc v_j est combinaison linéaire de $e_1, \ldots, e_n, f_1, \ldots, f_k$. Tous les vecteurs v_j sont donc combinaison de $e_1, \ldots, e_n, f_1, \ldots, f_k$ et comme ci-dessus, on en déduit que les vecteurs $e_1, \ldots, e_n, f_1, \ldots, f_k$ engendrent V. Comme ces vecteurs sont indépendants, ils forment une base de V. ■

Propriétés des bases. Nous les avons déjà énoncées dans le cadre de l'espace vectoriel $\mathbb{K}^p$; les démonstrations sont les mêmes, en utilisant une base quelconque de V au lieu de la base canonique de $\mathbb{K}^p$ (chapitre 4, paragraphe 3).

Soit V un $\mathbb{K}$-espace vectoriel engendré par un nombre fini de vecteurs.

1) Toutes les bases de V ont le même nombre d'éléments. Ce nombre s'appelle la *dimension* de V et se note $\dim V$.

2) Supposons que V est engendré par n vecteurs. Alors on a $\dim V \leqslant n$ et si $\dim V = n$, alors ces vecteurs forment une base de V.

3) Supposons que $u_1, u_2, \ldots, u_q$ sont des vecteurs indépendants appartenant à V. Alors on a $\dim V \geqslant q$ et si $q = \dim V$, alors ces vecteurs forment une base de V.

4) Si W est un sous-espace vectoriel de V, alors $\dim W \leqslant \dim V$ et l'on a l'équivalence $W = V \Longleftrightarrow \dim W = \dim V$.

Si un espace vectoriel est engendré par un nombre fini de vecteurs, on dit qu'il est *de dimension finie*.

Exemples.

- Les polynômes $1, z, \ldots, z^n$ forment une base de l'espace vectoriel $\mathbb{P}_n(\mathbb{K})$ des polynômes de degré inférieur ou égal à n : on a donc $\dim \mathbb{P}_n(\mathbb{K}) = n+1$.
- Les np matrices canoniques $E_{i,j}$ forment une base de l'espace vectoriel $\mathscr{M}_{p,n}(\mathbb{K})$, donc $\dim \mathscr{M}_{p,n}(\mathbb{K}) = np$.

1.2 Coordonnées d'un vecteur dans une base

Soient V un $\mathbb{K}$-espace vectoriel et $(e_1, e_2, \ldots, e_p)$ une base de V.
Tout vecteur $x \in V$ s'écrit de manière unique sous la forme $x = x_1e_1 + x_2e_2 + \cdots + x_pe_p$, où les x_i sont des scalaires de $\mathbb{K}$.

Définition

Les nombres $x_1, x_2, \ldots, x_p$ s'appellent les *coordonnées* du vecteur x dans la base $(e_1, e_2, \ldots, e_p)$.

À tout $x \in V$, associons le vecteur-colonne $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ formé des coordonnées de x dans la base $(e_1, e_2, \ldots, e_p)$. On a $X \in \mathbb{K}^p$.
Soient x et y des vecteurs appartenant à V.

- Si x a pour coordonnées X et si y a pour coordonnées Y, alors $x + y$ a pour coordonnées $X + Y$.
- Pour tout scalaire λ, le vecteur λx a pour coordonnées λX.

Les calculs dans l'espace vectoriel V se traduisent donc sur les coordonnées par les mêmes calculs dans $\mathbb{K}^p$. On en déduit les propriétés suivantes.

Propriété 1. Des vecteurs $u_1, u_2, \ldots, u_q$ de V sont indépendants si et seulement si leurs vecteurs de coordonnées $X_1, X_2, \ldots, X_q$ sont indépendants dans $\mathbb{K}^p$.

Propriété 2. Des vecteurs $u_1, u_2, \ldots, u_n$ engendrent V si et seulement si leurs vecteurs de coordonnées $X_1, X_2, \ldots, X_n$ engendrent $\mathbb{K}^p$.

Propriété 3. Soient $u_1, u_2, \ldots, u_p$ des vecteurs appartenant à V, de coordonnées $X_1, X_2, \ldots, X_p$. Les vecteurs $u_1, u_2, \ldots, u_p$ forment une base de V si et seulement si la matrice carrée de colonnes $X_1, X_2, \ldots, X_p$ est de rang p, c'est-à-dire si et seulement si $\det(X_1, X_2, \ldots, X_p) \neq 0$.

Changement de base

Supposons que $\mathscr{B} = (e_1, e_2, \dots, e_p)$ et $\mathscr{B}' = (e'_1, e'_2, \dots, e'_p)$ sont des bases de V.

Définition

Écrivons en colonne les coordonnées de $e'_1, e'_2, \dots, e'_p$ dans la base $\mathscr{B}$. On obtient une matrice carrée de taille p, appelée la *matrice de passage de la base $\mathscr{B}$ à la base $\mathscr{B}'$*.

Proposition. *Soient u un vecteur appartenant à V, $X = (x_1, x_2, \dots, x_p)$ ses coordonnées dans la base $\mathscr{B}$ et $X' = (x'_1, x'_2, \dots, x'_p)$ ses coordonnées dans la base $\mathscr{B}'$. On a $X = PX'$, où P est la matrice de passage de la base $\mathscr{B}$ à la base $\mathscr{B}'$.*

Démonstration. Dans la base $\mathscr{B}' = (e'_1, e'_2, \dots, e'_p)$, le vecteur-colonne des coordonnées de e'_i est $\boldsymbol{E_i}$, le i-ème vecteur canonique de $\mathbb{K}^p$. On sait que le produit $P\boldsymbol{E_i}$ est la i-ème colonne de P. Par définition de la matrice de passage, $P\boldsymbol{E_i}$ est donc le vecteur-colonne des coordonnées de e'_i dans la base $\mathscr{B}$. Puisque $u = x'_1 e'_1 + \dots + x'_p e'_p$, les coordonnées de u dans la base $\mathscr{B}$ sont $x'_1 P\boldsymbol{E_1} + \dots + x'_p P\boldsymbol{E_p} = P(x'_1\boldsymbol{E_1} + \dots + x'_p\boldsymbol{E_p}) = PX'$. ■

Une matrice de passage est toujours inversible : si P est la matrice de passage de $\mathscr{B}$ à $\mathscr{B}'$, alors P^{-1} est la matrice de passage de $\mathscr{B}'$ à $\mathscr{B}$.

Exemple. Les vecteurs $e'_1 = (1, -1, 2)$, $e'_2 = (1, 0, 1)$, $e'_3 = (2, 1, -1)$ forment une base de $\mathbb{R}^3$, car le déterminant $\begin{vmatrix} 1 & 1 & 2 \\ -1 & 0 & 1 \\ 2 & 1 & -1 \end{vmatrix} = -2$ est différent de 0. La matrice de passage de la base canonique de $\mathbb{R}^3$ à la base $\mathscr{B}' = (e'_1, e'_2, e'_3)$ étant $P = \begin{bmatrix} 1 & 1 & 2 \\ -1 & 0 & 1 \\ 2 & 1 & -1 \end{bmatrix}$, les coordonnées d'un vecteur $(x, y, z) \in \mathbb{R}^3$ dans la base $\mathscr{B}'$ sont données par le vecteur-colonne X' tel que $X = PX'$, c'est-à-dire $X' = P^{-1}X$. On a

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = P \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \iff \begin{cases} x' + y' + 2z' = x \\ -x' + z' = y \\ 2x' + y' - z' = z \end{cases} \iff \begin{cases} x' + y' + 2z' = x \\ y' + 3z' = x + y \\ -y' - 5z' = -2x + z \end{cases}$$

$$\iff \begin{cases} x' + y' + 2z' = x \\ y' + 3z' = x + y \\ 2z' = x - y - z \end{cases} \iff \begin{cases} 2x' = x - 3y - z \\ 2y' = -x + 5y + 3z \\ 2z' = x - y - z \end{cases}$$

On a donc $(x, y, z) = \frac{1}{2}(x - 3y - z)e'_1 + \frac{1}{2}(-x + 5y + 3z)e'_2 + \frac{1}{2}(x - y - z)e'_3$.

Remarque

Quand on connaît les coordonnées d'un vecteur dans une base, pour calculer ses coordonnées dans une nouvelle base, il faut inverser la matrice de passage.

2. Applications linéaires

Soit M une matrice à p lignes et n colonnes. Pour tout vecteur-colonne $X \in \mathbb{K}^n$, le produit MX est un vecteur-colonne de $\mathbb{K}^p$. D'après les propriétés du produit de matrices, on a les formules suivantes, pour tous vecteurs-colonne $X, X' \in \mathbb{K}^n$ et pour tout scalaire $\lambda \in \mathbb{K}$: $M(X+X') = MX + MX'$ et $M(\lambda X) = \lambda MX$. En notant $f : \mathbb{K}^n \to \mathbb{K}^p$ l'application $X \mapsto MX$, ces relations s'écrivent $f(X + X') = f(X) + f(X')$ et $f(\lambda X) = \lambda f(X)$.

Définition

Soient V et V' des $\mathbb{K}$-espaces vectoriels et soit $f : V \to V'$ une application. On dit que f est une *application linéaire* si pour tous vecteurs $u, v \in V$ et pour tout scalaire $\lambda \in \mathbb{K}$, on a $f(u + v) = f(u) + f(v)$ et $f(\lambda u) = \lambda f(u)$.

Propriétés des applications linéaires

1) Si f est une application linéaire de V dans V', alors pour toute combinaison linéaire de vecteurs de V, on a

$$f(\lambda_1 u_1 + \lambda_2 u_2 + \cdots + \lambda_k u_k) = \lambda_1 f(u_1) + \lambda_2 f(u_2) + \cdots + \lambda_k f(u_k)\,.$$

L'image par f d'une combinaison des u_i est la combinaison correspondante des $f(u_i)$.

2) Supposons que $\mathscr{B} = (e_1, \ldots, e_n)$ est une base de V. Une application linéaire $f : V \to V'$ est entièrement déterminée par les vecteurs $v'_1 = f(e_1), \ldots, v'_n = f(e_n)$ de V' : si un vecteur $x \in V$ a pour coordonnées $x_1, \ldots, x_n$ dans la base $\mathscr{B}$, on a en effet $f(x) = x_1 v'_1 + \cdots + x_n v'_n$.

3) Si $f : V \to V'$ est une application linéaire, alors $f(\mathbf{0}) = \mathbf{0}$.

En effet, si u est un vecteur quelconque de V, $0\,u = \mathbf{0}$ est le vecteur nul de V et $f(\mathbf{0}) = f(0\,u) = 0\,f(u) = \mathbf{0}$ est le vecteur nul de V'.

Une application linéaire de V dans lui-même s'appelle une *transformation linéaire de V*.

Exemples

- Si M est une matrice à p lignes et n colonnes, l'application $X \mapsto MX$ est une application linéaire de $\mathbb{K}^n$ dans $\mathbb{K}^p$.
- L'application qui à tout polynôme associe son polynôme dérivé est une application linéaire $\mathbb{P}_n(\mathbb{K}) \to \mathbb{P}_{n-1}(\mathbb{K})$. La multiplication par un polynôme A fixé est l'application linéaire $\mathbb{P}(\mathbb{K}) \to \mathbb{P}(\mathbb{K})$ définie par $P \mapsto AP$, pour tout $P \in \mathbb{P}(\mathbb{K})$.
- Donnons-nous une fonction continue $\omega : \mathbb{R} \to \mathbb{R}$ et pour toute fonction $t \mapsto y(t)$ deux fois dérivable sur $\mathbb{R}$, posons $D(y) = y'' + \omega y$, où ωy désigne la fonction produit $t \mapsto \omega(t)y(t)$. Si y et z sont des fonctions deux fois dérivables, on a $(y+z)'' = y'' + z''$ et $\omega(y+z) = \omega y + \omega z$, donc $D(y+z) = D(y) + D(z)$; de même, $D(\lambda y) = \lambda D(y)$ pour tout $\lambda \in \mathbb{R}$.
  En notant V l'espace vectoriel des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ et W le sous-espace vectoriel des fonctions deux fois dérivables, l'application $D : W \to V$ est donc linéaire.

► Soit V l'espace vectoriel des fonctions continues sur le segment $[a,b]$ et soit $\varphi : [a,b] \to \mathbb{R}$ une fonction donnée. D'après les propriétés de linéarité de l'intégrale, l'application $f \mapsto \int_a^b f(t)\varphi(t)\,dt$ est une application linéaire de V dans $\mathbb{R}$.

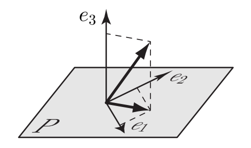
Une rotation. Posons $M_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. La transformation linéaire de $\mathbb{R}^2$ définie par $X \mapsto M_\theta X$ est la rotation de centre l'origine et d'angle θ.

Posons $\begin{bmatrix} x \\ y \end{bmatrix} = X$ et $\begin{bmatrix} x' \\ y' \end{bmatrix} = M_\theta X = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix}$. En introduisant les nombres complexes $z = x + \boldsymbol{i}\,y$ et $z' = x' + \boldsymbol{i}\,y'$, il vient en effet

$$z' = x(\cos\theta + \boldsymbol{i}\sin\theta) + \boldsymbol{i}\,y(\cos\theta + \boldsymbol{i}\sin\theta) = (x + \boldsymbol{i}\,y)e^{\boldsymbol{i}\,\theta} = ze^{\boldsymbol{i}\,\theta}\,.$$

On sait que cette formule traduit, sur les affixes des points, la rotation de centre l'origine et d'angle θ (page 41).

Une projection. Soit P un plan vectoriel de $\mathbb{R}^3$. Choisissons une base (e_1, e_2) de P et un vecteur e_3 n'appartenant pas à P.



Soit p la projection sur le plan P dans la direction e_3. On a $p(e_1) = e_1$, $p(e_2) = e_2$ et $p(e_3) = 0$. Puisque p est une transformation linéaire, cela détermine le projeté d'un vecteur quelconque : pour tout vecteur $xe_1 + ye_2 + ze_3$ de $\mathbb{R}^3$, on a en effet $p(xe_1 + ye_2 + ze_3) = xp(e_1) + yp(e_2) + zp(e_3) = xe_1 + ye_2$.

2.1 Matrice d'une application linéaire

Soient V et V' des $\mathbb{K}$-espaces vectoriels de dimension $n = \dim V$ et $p = \dim V'$. Donnons-nous une base $\mathscr{B} = (e_1, e_2, \ldots, e_n)$ de V et une base $\mathscr{B}' = (e'_1, e'_2, \ldots, e'_p)$ de V'. Soit $f : V \to V'$ une application linéaire.

Formons la matrice M dont la j-ème colonne est constituée des coordonnées de $f(e_j)$ dans la base $\mathscr{B}'$. Si $f(e_j) = a_{1j}e'_1 + a_{2j}e'_2 + \cdots + a_{pj}e'_p$ pour tout $j = 1, 2, \ldots, n$, on obtient

$$\begin{array}{cc} & \begin{array}{cccc} f(e_1) & f(e_2) & \cdots & f(e_n) \end{array} \\ M = & \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{bmatrix} \begin{array}{c} e'_1 \\ e'_2 \\ \vdots \\ e'_p \end{array} \end{array}$$

Définition

La matrice M s'appelle la *matrice de l'application linéaire* f dans les bases $\mathscr{B}$ et $\mathscr{B}'$. Si $V = V'$ et $\mathscr{B} = \mathscr{B}'$, on dit simplement que M est la matrice de f dans la base $\mathscr{B}$.

Soit v un vecteur appartenant à V, de coordonnées $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ dans la base $\mathscr{B}$. Puisque $v = x_1e_1 + \cdots + x_ne_n$, on a l'égalité $f(v) = x_1f(e_1) + x_2f(e_2) + \cdots + x_nf(e_n)$ dans l'espace vectoriel V'. Prenons les coordonnées de chaque membre dans la base $\mathscr{B}'$, en notant Y le

vecteur-colonne des coordonnées de $f(v)$. Par définition de la matrice M de f, les coordonnées de $f(e_j)$ forment la j-ième colonne de M ; en appelant M_j cette colonne, il vient

$$\begin{aligned} Y &= x_1 M_1 + x_2 M_2 \cdots + x_n M_n = x_1 M\boldsymbol{E_1} + x_2 M\boldsymbol{E_2} + \cdots + x_n M\boldsymbol{E_n} \ , \quad \text{car } M_j = M\boldsymbol{E_j} \\ &= M(x_1\boldsymbol{E_1} + x_2\boldsymbol{E_2} + \cdots + x_n\boldsymbol{E_n}) = MX \,. \end{aligned}$$

Proposition. *Soit M la matrice d'une application linéaire $f : V \to V'$ dans des bases $\mathscr{B}$ et $\mathscr{B}'$. Si le vecteur $v \in V$ a pour coordonnées X dans la base $\mathscr{B}$, alors $f(v)$ a pour coordonnées MX dans la base $\mathscr{B}'$.*

Sur les coordonnées des vecteurs, l'application f se traduit par l'application $X \mapsto MX$ de $\mathbb{K}^n$ dans $\mathbb{K}^p$.

Soit f une transformation linéaire de V et soit $\varphi : V \to \mathbb{K}^n$ la bijection qui à tout vecteur $v \in V$ associe ses coordonnées X dans une base $\mathscr{B}$. Notons $L : \mathbb{K}^n \to \mathbb{K}^n$ la transformation $X \mapsto MX$, où M est la matrice de f dans $\mathscr{B}$. On obtient L à partir de f par le changement de référentiel φ (page 24), autrement dit : $\varphi \circ f \circ \varphi^{-1} = L$.

$$\begin{array}{ccc} V & \xrightarrow{f} & V \\ \varphi \big| & & \big| \varphi \\ \mathbb{K}^n & \xrightarrow[L]{} & \mathbb{K}^n \end{array} \qquad\qquad \begin{array}{ccc} v & \xrightarrow{f} & f(v) \\ \varphi \big| & & \big| \varphi \\ X & \longrightarrow & MX \end{array}$$

Exemple. Soit a un nombre réel. Pour tout polynôme $P(z)$, posons $f(P) = zP'' + aP$. Si P est de degré n, alors zP'' est de degré $n-2+1 = n-1$, donc $f(P)$ est de degré n si $a \neq 0$, de degré $n-1$ si $a = 0$. On définit ainsi une application $f : \mathbb{P}_n \to \mathbb{P}_n$ et comme la dérivation est une opération linéaire, f est linéaire.
Supposons $n = 3$ et calculons la matrice de f dans la base $(1, z, z^2, z^3)$ de l'espace vectoriel $\mathbb{P}_3$. On a $f(1) = a$, $f(z) = az$, $f(z^2) = 2z + az^2$ et $f(z^3) = 6z^2 + az^3$.

La matrice de f dans la base $(1, z, z^2, z^3)$ est donc $\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 2 & 0 \\ 0 & 0 & a & 6 \\ 0 & 0 & 0 & a \end{bmatrix}$.

Remarques

Si V est un $\mathbb{K}$-espace vectoriel de dimension n, alors

- toute transformation linéaire de V a une matrice carrée de taille n ;
- dans n'importe quelle base de V, la matrice de la transformation identité est la matrice unité I_n.

2.2 Calcul sur les applications linéaires

Soient V et V' des $\mathbb{K}$-espaces vectoriels. Les propriétés suivantes sont immédiates.

- Si f et g sont des applications linéaires de V dans V', la somme $f+g : v \mapsto f(v)+g(v)$ est une application linéaire de V dans V', de même que l'application $\lambda f : v \mapsto \lambda f(v)$.
- La composée de deux applications linéaires est linéaire.

► Si f est une application linéaire bijective, la bijection réciproque est une application linéaire.

Supposons maintenant que $\mathscr{B}$ est une base de V et $\mathscr{B}'$ une base de V'.

Proposition. *Soient f et g des applications linéaires de V dans V', de matrices M et N dans les bases $\mathscr{B}$ et $\mathscr{B}'$.*

i) *L'application linéaire $f+g$ a pour matrice $M+N$ et λf a pour matrice λM.*

ii) *Supposons $\dim V = \dim V'$. L'application f est bijective si et seulement si la matrice M est inversible et dans ce cas, la matrice de f^{-1} dans les bases $\mathscr{B}'$ et $\mathscr{B}$ est M^{-1}.*

Démonstration. La propriété (i) est évidente. Supposons $\dim V = \dim V' = n$. Pour tous vecteurs $v \in V$ et $v' \in V'$, la relation $v' = f(v)$ équivaut à $X' = MX$, où X est le vecteur-colonne des coordonnées de v dans $\mathscr{B}$ et X' celui des coordonnées de v' dans $\mathscr{B}'$. L'application f est donc bijective si et seulement si l'application $X \mapsto MX$ de $\mathbb{K}^n$ dans $\mathbb{K}^n$ est bijective. D'après la proposition page 138, on en déduit que f est bijective si et seulement si M est inversible. Dans ce cas, on a les équivalences $v = f^{-1}(v') \Longleftrightarrow v' = f(v) \Longleftrightarrow X' = MX \Longleftrightarrow X = M^{-1}X'$, donc la matrice de f^{-1} dans les bases $\mathscr{B}'$ et $\mathscr{B}$ est M^{-1}. ■

Proposition. *Soient $f, g : V \to V$ des transformations linéaires de V. Si f et g ont pour matrice M et N dans la base $\mathscr{B}$, alors $f \circ g$ a pour matrice MN dans la base $\mathscr{B}$.*

Démonstration. Posons $n = \dim V$. Soient $v \in V$ et $X \in \mathbb{K}^n$ le vecteur-colonne des coordonnées de v dans la base $\mathscr{B}$. Les coordonnées de $g(v)$ sont données par le vecteur-colonne $Y = NX$ et les coordonnées de $f \circ g(v) = f\big(g(v)\big)$ sont données par $Z = MY$. Puisque $Z = M(NX) = (MN)X$, la matrice de $g \circ f$ dans la base $\mathscr{B}$ est MN. ■

Exemples

► Dans $\mathbb{R}^2$, la rotation de centre l'origine et d'angle θ est $X \mapsto M(\theta)X$, où $M(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. Si l'on compose cette rotation avec une rotation de même centre et d'angle θ', on obtient la rotation d'angle $\theta' + \theta$. Par conséquent $M(\theta')M(\theta) = M(\theta' + \theta)$. On en déduit que pour tout entier $n \geqslant 1$, on a

$$\big(M(\theta)\big)^n = M(n\theta) \quad \text{et} \quad \big(M(\theta)\big)^{-1} = M(-\theta) = {}^t\big(M(\theta)\big).$$

► Soit P un plan vectoriel de $\mathbb{R}^3$, de base (e_1, e_2), et soit e_3 un vecteur orthogonal à P : si P a pour équation $ax+by+cz=0$, le vecteur $e_3 = (a,b,c)$ convient. Notons $p : \mathbb{R}^3 \to \mathbb{R}^3$ la projection orthogonale sur P, donc $p(xe_1 + ye_2 + ze_3) = xe_1 + ye_2$ (figure page 168). La matrice de p dans la base (e_1, e_2, e_3) est $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Puisque $p \circ p = p$, on en déduit $M^2 = M$.

Changement de base

Soient $\mathscr{B}$ et $\mathscr{B}'$ des bases du même espace vectoriel V de dimension n et soit $P \in \mathscr{M}_n(\mathbb{K})$ la matrice de passage de $\mathscr{B}$ à $\mathscr{B}'$ (définition page 166).

Formule de changement de base. *Soit $f: V \to V$ une transformation linéaire. Si M est la matrice de f dans la base $\mathscr{B}$, la matrice de f dans la base $\mathscr{B}'$ est $P^{-1}MP$.*

Démonstration. Tout vecteur $v \in V$ a des coordonnées X dans la base $\mathscr{B}$, des coordonnées X' dans la base $\mathscr{B}'$ et l'on a $X = PX'$. Notons Y les coordonnées de $f(v)$ dans la base $\mathscr{B}$ et Y' les coordonnées de $f(v)$ dans $\mathscr{B}'$ de sorte qu'on a aussi $Y = PY'$. Puisque M est la matrice de f dans la base $\mathscr{B}$, on a $Y = MX$ et il vient $Y' = P^{-1}Y = P^{-1}(MX) = (P^{-1}M)(PX') = (P^{-1}MP)X'$. La matrice de f dans la base $\mathscr{B}'$ est donc $P^{-1}MP$. ■

Un changement de base s'exprime par un changement de coordonnées (page 24) :

par le changement de coordonnées $X' = P^{-1}X$, la transformation $X \mapsto MX$ de $\mathbb{K}^n$ devient $X' \mapsto (P^{-1}MP)X'$.

Polynôme caractéristique d'une transformation linéaire

Soit $f: V \to V$ une transformation linéaire, de matrices M et M' dans les bases $\mathscr{B}$ et $\mathscr{B}'$. Rappelons que nous avons défini le polynôme caractéristique de M (page 151) : c'est le polynôme $C_M(z) = \det(M - zI_n)$. Puisque $M' = P^{-1}MP$, où P est la matrice de passage de $\mathscr{B}$ à $\mathscr{B}'$, on a $M' - zI_n = P^{-1}MP - P^{-1}zI_nP = P^{-1}(M - zI_n)P$. Comme le déterminant d'un produit de matrices est le produit des déterminants, on en déduit

$$\begin{aligned} C_{M'} &= \det(M' - zI_n) = \det\big[P^{-1}(M - zI_n)P\big] = \det(P^{-1})\det(M - zI_n)\det P \\ &= \det(M - zI_n)\det(P^{-1})\det P = C_M(z)\ , \quad \text{car } \det(P^{-1}) = (\det P)^{-1}. \end{aligned}$$

Si des matrices représentent la même transformation linéaire dans des bases différentes, elles ont le même polynôme caractéristique.

Le polynôme C_M ne dépend donc que de la transformation f.

Définition

Si la transformation f a pour matrice M dans une base de V, le polynôme C_M s'appelle le *polynôme caractéristique de f* et se note C_f.

Pour calculer le polynôme caractéristique d'une transformation f, on peut utiliser la matrice de f dans n'importe quelle base.

Conséquence. Comme le polynôme caractéristique ne dépend que de f, chacun de ses coefficients ne dépend que de f. Puisqu'on a

$$C_M(z) = (-1)^n z^n + (-1)^{n-1}(\operatorname{tr} M)z^{n-1} + \cdots + a_1 z + \det M\ ,$$

on en déduit que

- le déterminant de M ne dépend que de f : on pose $\det f = \det M$.
- la trace de M ne dépend que de f : on pose $\operatorname{tr} f = \operatorname{tr} M$.

2.3 Noyau et image d'une application linéaire

Soient V et V' des $\mathbb{K}$-espaces vectoriels et $f : V \to V'$ une application linéaire.

► L'ensemble $N = \{u \in V \mid f(u) = \mathbf{0}\}$ est un sous-espace vectoriel de V.

> Le vecteur nul de V appartient à N, car $f(\mathbf{0}) = \mathbf{0}$. De plus, si u et v appartiennent à N, on a $f(u+v) = f(u) + f(v) = \mathbf{0} + \mathbf{0} = \mathbf{0}$ et pour tout $\lambda \in \mathbb{K}$, $f(\lambda u) = \lambda f(u) = \lambda \mathbf{0} = \mathbf{0}$, donc $u + v$ et λu appartiennent à N.

► Soit W un sous-espace vectoriel de V. L'ensemble $f(W)$ des vecteurs $f(w)$, où w parcourt W, est un sous-espace vectoriel de V'.

> En effet, si v'_1 et v'_2 sont des vecteurs de la forme $v'_1 = f(w_1)$ et $v'_2 = f(w_2)$, alors $w_1 + w_2 \in W$ et $v'_1 + v'_2 = f(w_1 + w_2)$ est bien image par f d'un vecteur de W ; de même, λw_1 appartient à W et $\lambda v'_1 = f(\lambda w_1)$.

Définitions

Soit $f : V \to V'$ une application linéaire.

► Le sous-espace vectoriel de V formé des vecteurs $v \in V$ tels que $f(v) = \mathbf{0}$ s'appelle le *noyau* de f et se note $\operatorname{Ker} f$.

► L'image de l'application f, c'est-à-dire l'ensemble $f(V)$, est un sous-espace vectoriel de V' noté $\operatorname{Im} f$.

Lorsque les espaces vectoriels sont de dimension finie, le calcul matriciel permet de trouver les vecteurs du noyau et de l'image : on utilise pour cela la proposition suivante.

Proposition. *Supposons que $\mathscr{B}$ et $\mathscr{B}'$ sont des bases de V et V'. Soit $f : V \to V'$ une application linéaire et soit M la matrice de f dans ces bases.*

► *Un vecteur $v \in V$ est dans $\operatorname{Ker} f$ si et seulement si ses coordonnées X sont solution du système linéaire $MX = 0$.*

► *Un vecteur $v' \in V'$ est dans $\operatorname{Im} f$ si et seulement si ses coordonnées sont combinaison linéaire des vecteurs-colonne de M.*

Corollaire. *Supposons V et V' de dimension finie. On a les égalités :*

i) $\dim \operatorname{Ker} f = \dim V - \operatorname{rg} M$;

ii) $\dim \operatorname{Im} f = \operatorname{rg} M$;

iii) $\dim \operatorname{Ker} f + \dim \operatorname{Im} f = \dim V$ *(formule de la dimension)*

Démonstration. Les solutions de l'équation $MX = 0$ forment un sous-espace vectoriel de dimension $n - r$, où $n = \dim V$ et $r = \operatorname{rg} M$ (résultat 2 page 114). D'après les propriétés page 165, ce sous-espace vectoriel a la même dimension que le noyau de f, d'où (i). Posons $\dim V' = p$. Par définition de la matrice de f, la dimension de $\operatorname{Im} f$ est celle du sous-espace vectoriel de $\mathbb{K}^p$ engendré par les colonnes de M. D'après le théorème page 137, cette dimension est égale au rang de M, ce qui montre (ii). L'égalité (iii) résulte immédiatement de (i) et (ii). ■

Corollaire. *Soient V et V' des espaces vectoriels tels que* $\dim V = \dim V'$ *et soit* $f : V \to V'$ *une application linéaire.*

- *f est bijective si et seulement si* $\operatorname{Ker} f = \{\mathbf{0}\}$.
- *On a les équivalences : f est bijective* $\iff \operatorname{Im} f = V' \iff \dim(\operatorname{Im} f) = \dim V'$.

Démonstration. Choisissons des bases de V et V' et soit M la matrice de f. Puisque $\dim V = \dim V' = n$, la matrice M est carrée de taille n. D'après une proposition page 170, l'application linéaire f est bijective si et seulement si M est inversible. Utilisons la proposition page 138 : M est inversible si et seulement si le système linéaire $MX = 0$ a pour seule solution $X = 0$, ce qui équivaut à $\operatorname{Ker} f = \{0\}$. D'après la même proposition, M est inversible si et seulement si $\operatorname{rg} M = n$, ce qui équivaut à $\dim(\operatorname{Im} f) = \dim V'$, ou encore à $\operatorname{Im} f = V'$ puisque $\operatorname{Im} f$ est un sous-espace vectoriel de V' (propriété (4) page 165). ■

3. Diagonalisation

Dans ce paragraphe,

- V est un $\mathbb{K}$-espace vectoriel de dimension n et $(e_1, e_2, \ldots, e_n)$ est une base de V ;
- f est une transformation linéaire de V et M est la matrice de f dans la base $(e_1, e_2, \ldots, e_n)$.

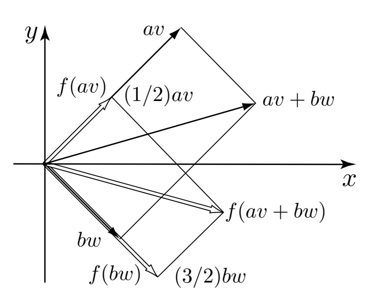
Définitions

Un vecteur v non nul tel que $f(v)$ est colinéaire à v s'appelle un *vecteur propre* de f. Si v est un vecteur propre, le scalaire $\lambda \in \mathbb{K}$ tel que $f(v) = \lambda v$ s'appelle la *valeur propre* associée. On dit aussi que v est vecteur propre pour la valeur propre λ.

Exemple. Pour la transformation de $\mathbb{R}^2$ définie par $f(x, y) = (x - y/2, -x/2 + y)$:

- on a $f(1,1) = (1/2)(1,1)$, donc le vecteur $v = (1,1)$ est vecteur propre pour la valeur propre $1/2$;
- le vecteur $w = (1, -1)$ est propre pour la valeur propre $3/2$, car $f(w) = (3/2)w$.

La droite dirigée par v est transformée en elle-même, de même que la droite dirigée par w.



Exemples

- Si $r : \mathbb{R}^2 \to \mathbb{R}^2$ est une rotation de centre l'origine et d'angle $\theta \in [0, 2\pi[$ et si v est un vecteur non nul, l'angle $\widehat{v, r(v)}$ a pour mesure θ : une rotation d'angle différent de 0 et de π n'a donc pas de vecteur propre.
- Soit $p : \mathbb{R}^3 \to \mathbb{R}^3$ la projection orthogonale sur un plan vectoriel P (figure page 168). Pour tout vecteur $v \in P$, on a $p(v) = v$: tout vecteur non nul du plan P est donc vecteur propre de p, avec 1 comme valeur propre. Soit u un vecteur non

nul et orthogonal à P ; alors $p(u)$ est le vecteur nul, donc $p(u) = 0u$: le vecteur u est vecteur propre pour la valeur propre 0.

- D'une manière générale, les éventuels vecteurs non nuls de $\operatorname{Ker} f$ sont les vecteurs propres de f pour la valeur propre 0.

Supposons que λ est une valeur propre de f. Pour tout $x \in V$, posons $g(x) = f(x) - \lambda x$, ce qui définit la transformation linéaire $g = f - \lambda\, \mathrm{id}_V$. Par hypothèse, il existe un vecteur propre v tel que $f(v) = \lambda v$, donc $g(v) = \mathbf{0}$. Le vecteur v est non nul et appartient au noyau de g : d'après le dernier corollaire du paragraphe précédent, l'application g n'est donc pas bijective.

Réciproquement, si λ est un scalaire et si la transformation $f - \lambda\, \mathrm{id}_V$ n'est pas bijective, il existe un vecteur v non nul dans le noyau de $f - \lambda\, \mathrm{id}_V$. On a alors $\mathbf{0} = (f - \lambda\, \mathrm{id}_V)(v) = f(v) - \lambda v$, donc $f(v) = \lambda v$ et λ est une valeur propre. Puisque la matrice de $f - \lambda\, \mathrm{id}_V$ est $M - \lambda I_n$, on en déduit :

un scalaire $\lambda \in \mathbb{K}$ est une valeur propre de f si et seulement si la matrice $M - \lambda I_n$ n'est pas inversible.

Rappelons que, par définition, le polynôme caractéristique de f est $C_f(z) = \det(M - zI_n)$ (page 171). Puisque la matrice $M - \lambda I_n$ est inversible si et seulement si son déterminant est non nul, on a la proposition suivante.

Proposition. *Les valeurs propres de f sont les nombres $\lambda \in \mathbb{K}$ qui sont racines du polynôme caractéristique de f.*

Le polynôme caractéristique de f étant de degré n, f possède au plus n valeurs propres.

Remarques

- Si M est une matrice triangulaire, les valeurs propres de f sont les nombres situés sur la diagonale de M.

  En effet, si $M = [m_{ij}]$ est triangulaire, alors $M - zI_n$ aussi, donc $\det(M - zI_n)$ est le produit $(m_{11} - z)(m_{22} - z) \cdots (m_{nn} - z)$ des termes diagonaux.

- Si t est un scalaire, la matrice $M - tI_n$ est inversible sauf pour un nombre fini de valeurs exceptionnelles de t, constituées précisément des racines du polynôme caractéristique de M.
- Si toutes les racines du polynôme caractéristique de M sont dans $\mathbb{K}$, la somme des valeurs propres de M est égale à la trace de M.

  Si ces racines sont $\lambda_1, \ldots, \lambda_n$, alors $(-1)^n C_M = (z - \lambda_1) \cdots (z - \lambda_n)$ (page 171). Puisque le coefficient de z^{n-1} est $-\operatorname{tr} M$, il vient $-\operatorname{tr} M = -\lambda_1 - \lambda_2 - \cdots - \lambda_n$.

Exemple 1. Reprenons la transformation f de $\mathbb{R}^2$ de matrice $M = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}$ (exemple page 173). On a

$$\det(M - zI_2) = \begin{vmatrix} 1-z & -1/2 \\ -1/2 & 1-z \end{vmatrix} = z^2 - 2z + \frac{3}{4} = \left(z - \frac{1}{2}\right)\left(z - \frac{3}{2}\right)$$

Les valeurs propres de f sont $1/2$ et $3/2$. Pour calculer, par exemple, les vecteurs propres pour la valeur propre $1/2$, on résout le système linéaire $\left(M - \frac{1}{2}I_2\right)X = 0$: les solutions sont les vecteurs $t(1,1)$.

Exemple 2. Posons $M_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, où $0 \leqslant \theta < 2\pi$. Le polynôme caractéristique de M est

$$C_M = \det(A - zI_2) = \begin{vmatrix} \cos\theta - z & -\sin\theta \\ \sin\theta & \cos\theta - z \end{vmatrix} = (\cos\theta - z)^2 + \sin^2\theta\,.$$

- Supposons θ différent de 0 et de π, c'est-à-dire $\sin\theta \neq 0$. Alors C_M n'a pas de racine réelle, ce qui correspond au fait que la rotation d'angle θ n'a pas de vecteur propre.
- Supposons $\theta = 0$. On a $M_0 = I_2$, le polynôme caractéristique est $(1-z)^2$ et la transformation de $\mathbb{R}^2$ définie par $X \mapsto M_0 X$ est l'identité : tout vecteur non nul de $\mathbb{R}^2$ est propre pour la valeur propre 1.
- Supposons $\theta = \pi$. On a $M_\pi = -I_2$, le polynôme caractéristique est $(1+z)^2$ et la transformation de $\mathbb{R}^2$ définie par $X \mapsto M_\pi X$ est $-\mathrm{id}_{\mathbb{R}^2}$: c'est la symétrie par rapport à l'origine. Tout vecteur non nul de $\mathbb{R}^2$ est propre pour la valeur propre -1.
- Soit $f : \mathbb{C}^2 \to \mathbb{C}^2$ la transformation définie par $X \mapsto M_\theta X$. Puisqu'on a $C_M = z^2 - (2\cos\theta)z + 1 = (z - e^{\boldsymbol{i}\theta})(z - e^{-\boldsymbol{i}\theta})$, les valeurs propres de f sont $e^{\boldsymbol{i}\theta}$ et $e^{-\boldsymbol{i}\theta}$. On a

  $$M_\theta \begin{bmatrix} 1 \\ \boldsymbol{i} \end{bmatrix} = \begin{bmatrix} \cos\theta - \boldsymbol{i}\sin\theta \\ \sin\theta + \boldsymbol{i}\cos\theta \end{bmatrix} = e^{-\boldsymbol{i}\theta} \begin{bmatrix} 1 \\ \boldsymbol{i} \end{bmatrix} \quad \text{et} \quad M_\theta \begin{bmatrix} 1 \\ -\boldsymbol{i} \end{bmatrix} = e^{\boldsymbol{i}\theta} \begin{bmatrix} 1 \\ -\boldsymbol{i} \end{bmatrix}$$

  donc $\begin{bmatrix} 1 \\ -\boldsymbol{i} \end{bmatrix}$ est vecteur propre de f pour la valeur propre $e^{\boldsymbol{i}\theta}$ et le vecteur conjugué $\begin{bmatrix} 1 \\ \boldsymbol{i} \end{bmatrix}$ est propre pour la valeur propre conjuguée $e^{-\boldsymbol{i}\theta}$.

Étude des vecteurs propres

Proposition. *Supposons que $\lambda_1, \lambda_2, \ldots, \lambda_k$ sont des valeurs propres deux à deux différentes. Si $u_1, u_2, \ldots, u_k$ sont des vecteurs propres pour ces valeurs propres, alors $u_1, u_2, \ldots, u_k$ sont indépendants.*

Démonstration. Si $k = 1$, la propriété est vraie car u_1 est un vecteur non nul. Raisonnons par récurrence en supposant que la propriété est vraie pour $k-1$ vecteurs. Supposons que les vecteurs $u_1, u_2, \ldots, u_k$ satisfont la relation

(1) $$x_1u_1 + x_2u_2 + \cdots + x_ku_k = \mathbf{0}\,.$$

Par hypothèse, on a $f(u_i) = \lambda_i u_i$, donc en appliquant f à l'égalité (1), il vient

(2) $$x_1\lambda_1u_1 + x_2\lambda_2u_2 + \cdots + x_k\lambda_ku_k = \mathbf{0}\,.$$

Multiplions (1) par λ_k et soustrayons à (2) :

$$x_1(\lambda_1-\lambda_k)u_1+x_2(\lambda_2-\lambda_k)u_2+\cdots+x_{k-1}(\lambda_{k-1}-\lambda_k)u_{k-1}=\mathbf{0}\,.$$

Les vecteurs $u_1,\ldots,u_{k-1}$ sont indépendants. D'après l'hypothèse de récurrence, on a donc $x_i(\lambda_i-\lambda_k)=0$ pour tout $i=1,\ldots,k-1$. Puisque $\lambda_i\neq\lambda_k$, on en déduit $x_i=0$ pour tout $i=1,\ldots,k-1$. L'égalité (1) devient $x_ku_k=0$, d'où $x_k=0$, car u_k n'est pas le vecteur nul. ■

Définition

Soit λ une valeur propre de f. Le noyau de la transformation $f-\lambda\,\mathrm{id}_V$ s'appelle le *sous-espace propre pour la valeur propre* λ et se note $V(\lambda)$.

Un vecteur v est dans $V(\lambda)$ si et seulement si $f(v)-\lambda v=\mathbf{0}$: les vecteurs propres de f pour la valeur propre λ sont exactement les vecteurs non nuls appartenant à $V(\lambda)$. Un sous-espace propre contient donc des vecteurs non nuls.

Notation. Si λ est une valeur propre de f, notons $m(\lambda)$ la multiplicité de λ comme racine du polynôme caractéristique de f (page 47). Si λ est racine simple du polynôme caractéristique, on dit que c'est une *valeur propre simple*.

Proposition. *Pour toute valeur propre λ de f, on a $\dim V(\lambda)\leqslant m(\lambda)$.*

Démonstration. Soient $v_1,v_2,\ldots,v_d$ une base de $V(\lambda)$. Si $d=n$, alors $V(\lambda)=V$: dans ce cas, tous les vecteurs non nuls de V sont propres avec λ pour valeur propre ; on a alors $f=\lambda\,\mathrm{id}_V$, $C_f(z)=(\lambda-z)^n$ et la multiplicité de λ est n. Supposons maintenant $d<n$. D'après le théorème de la base incomplète (page 164), il existe une base de V de la forme $(v_1,\ldots,v_d,w_{d+1},\ldots,w_n)$. Puisqu'on a $f(v_i)=\lambda v_i$ pour $i=1,\ldots,d$, la matrice de f dans cette base est de la forme

$$M'=\begin{bmatrix}\boxed{\lambda I_d} & *\\ 0 & \boxed{N}\end{bmatrix},\quad \text{où } N \text{ est une matrice carrée de taille } n-d.$$

On a $M'-zI_n=\begin{bmatrix}\boxed{(\lambda-z)I_d} & *\\ 0 & \boxed{N-zI_{n-d}}\end{bmatrix}$ et d'après la proposition page 151, il vient

$$C_f(z)=\det(M'-zI_n)=\det\big[(\lambda-z)I_d\big]\det(N-zI_{n-d})=(\lambda-z)^d\det(N-zI_{n-d})\,.$$

La racine λ de C_f a donc une multiplicité au moins égale à d. ■

Transformation diagonalisable

Proposition. *Pour que la matrice de f dans une base soit diagonale, il faut et il suffit que les vecteurs de cette base soient tous propres. Dans ce cas, le j-ème coefficient diagonal est la valeur propre associée au j-ème vecteur de base.*

Démonstration. Supposons que, dans la base $(u_1,u_2,\ldots,u_n)$, la matrice de f est la matrice diagonale $\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$. Puisque la j-ème colonne de cette matrice est $\lambda_j\boldsymbol{E_j}$, on a $f(u_j)=\lambda_ju_j$, donc les vecteurs de la base sont propres. Réciproquement, supposons que

les vecteurs de base $u_1, u_2, \ldots, u_n$ sont propres, avec valeurs propres associées $\lambda_1, \lambda_2, \ldots, \lambda_n$. On a $f(u_j) = \lambda_j u_j$, donc la j-ème colonne de la matrice de f dans cette base est $\lambda_j \boldsymbol{E_j}$: la matrice de f est donc $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. ■

Définition

La transformation linéaire f est *diagonalisable* s'il existe une base de V dans laquelle la matrice de f est diagonale. Une matrice $M \in \mathscr{M}_n(\mathbb{K})$ est dite diagonalisable si la transformation $X \mapsto MX$ de $\mathbb{K}^n$ est diagonalisable.

Une transformation linéaire f est diagonalisable si et seulement s'il existe une base de V formée de vecteurs propres de f.

Exemple. Soient $M = \begin{bmatrix} 1 & 1 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix}$ et $f : \mathbb{R}^3 \to \mathbb{R}^3$ la transformation $X \mapsto MX$.
Le polynôme caractéristique de f est $C_f = \begin{vmatrix} 1-z & 1 & 3 \\ 0 & 2-z & 3 \\ 0 & 0 & 1-z \end{vmatrix} = (1-z)^2(2-z)$. Il y a deux valeurs propres : les nombres 1 et 2.

i) Puisque la première colonne de M est $\boldsymbol{E_1}$, on a $M\boldsymbol{E_1} = \boldsymbol{E_1}$ donc le vecteur $\boldsymbol{E_1}$ est propre pour la valeur propre 1.

ii) On a $M - I_3 = \begin{bmatrix} 0 & 1 & 3 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}$ et le vecteur $U_2 = \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}$ est solution du système linéaire $(M - I_3)X = 0$. On a $MU_2 = U_2$, donc U_2 est propre pour la valeur propre 1.

iii) On a $M - 2I_3 = \begin{bmatrix} -1 & 1 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & -1 \end{bmatrix}$ et le vecteur $U_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ est solution du système linéaire $(M - 2I_3)X = 0$. On a $MU_3 = 2U_3$, donc U_3 est propre pour la valeur propre 2.

Les vecteurs $\boldsymbol{E_1}, U_2, U_3$ étant indépendants, ils forment une base de $\mathbb{R}^3$. La transformation f est donc diagonalisable et dans la base (E_1, U_2, U_3), la matrice de f est $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \operatorname{diag}(1, 1, 2)$.

Critère de diagonalisation. *Pour que f soit diagonalisable, il faut et il suffit que le polynôme caractéristique de f ait toutes ses racines dans $\mathbb{K}$ et que pour toute racine λ de C_f, on ait $\dim V(\lambda) = m(\lambda)$.*

Démonstration. Supposons que le polynôme caractéristique a toutes ses racines $\lambda_1, \ldots, \lambda_k$ dans $\mathbb{K}$, donc $C_f = (\lambda_1 - z)^{m(\lambda_1)}(\lambda_2 - z)^{m(\lambda_2)} \cdots (\lambda_k - z)^{m(\lambda_k)}$. Choisissons une base $\mathscr{B}_i$ du sous-espace propre $V(\lambda_i)$ et soit $\mathscr{B}$ la réunion de ces bases. Une combinaison linéaire de vecteurs de $\mathscr{B}$ est de la forme $v_1 + v_2 + \cdots + v_k$, où v_i est combinaison des vecteurs de $\mathscr{B}_i$, donc $v_i \in V(\lambda_i)$. Supposons $v_1 + v_2 + \cdots + v_k = \mathbf{0}$. Puisqu'un vecteur non nul de $V(\lambda_i)$ est un vecteur propre pour λ_i, on en déduit que tous les vecteurs v_i sont nuls, en utilisant la proposition page 175. Il s'ensuit que dans l'expression de v_i comme combinaison des vecteurs de $\mathscr{B}_i$, tous les coefficients sont nuls, car les vecteurs de $\mathscr{B}_i$ sont indépendants. Ainsi les vecteurs de $\mathscr{B}$ sont indépendants.

Après ce préliminaire, montrons que la condition est suffisante. Supposons $\lambda_i \in \mathbb{K}$ et $\dim V(\lambda_i) = m(\lambda_i)$ pour tout i. Chaque base $\mathscr{B}_i$ possède $m(\lambda_i)$ vecteurs, donc le nombre de vecteurs de $\mathscr{B}$ est $m(\lambda_1) + \cdots + m(\lambda_k) = \deg C_f = n$. Puisque les vecteurs de $\mathscr{B}$ sont indépendants, $\mathscr{B}$ est une base de V formée de vecteurs propres.
Montrons que la condition est nécessaire. Supposons f diagonalisable. La matrice de f dans une base $\mathscr{B}$ de vecteurs propres est de la forme $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, où $d_i \in \mathbb{K}$. Comme on a $C_f = \det(D - zI_n) = (d_1 - z)(d_2 - z) \cdots (d_n - z)$, toutes les racines de C_f sont dans $\mathbb{K}$, donc sont des valeurs propres. Comme dans le préliminaire, appelons $\lambda_1, \ldots, \lambda_k$ les différentes valeurs propres. Dans la base $\mathscr{B}$, regroupons les vecteurs propres relatifs à λ_i et appelons $\mathscr{B}_i$ la partie obtenue. En notant n_i le nombre d'éléments de $\mathscr{B}_i$, on a donc $n_1 + n_2 + \cdots + n_k = n$, le nombre d'éléments de $\mathscr{B}$. Les vecteurs de $\mathscr{B}_i$ sont indépendants et sont propres pour λ_i, donc $n_i \leqslant \dim V(\lambda_i)$. Puisqu'on a $\dim V(\lambda_i) \leqslant m(\lambda_i)$ d'après une précédente proposition, il vient

$$n = n_1 + n_2 + \cdots + n_k \leqslant m(\lambda_1) + m(\lambda_2) + \cdots + m(\lambda_k) = n$$

ce qui implique $n_i = m(\lambda_i)$ pour tout i. Il s'ensuit $\dim V(\lambda_i) = m(\lambda_i)$. ■

Corollaire. *Si f possède n valeurs valeurs propres distinctes, alors f est diagonalisable.*

Démonstration. L'hypothèse signifie que le polynôme caractéristique a n racines distinctes appartenant à $\mathbb{K}$, donc ces racines sont simples : pour toute valeur propre λ, on a $m(\lambda) = 1$, donc aussi $\dim V(\lambda) \leqslant 1$. Comme un sous-espace propre n'est pas nul, on en déduit que $V(\lambda)$ est de dimension 1. D'après le critère de diagonalisation, f est diagonalisable. ■

Exemple. Soient $a \in \mathbb{R}$, $M = \begin{bmatrix} 1 & a \\ -a & 2 \end{bmatrix}$ et $f : \mathbb{R}^2 \to \mathbb{R}^2$ la transformation $X \mapsto MX$. Le polynôme caractéristique de M est

$$C_M = \begin{vmatrix} 1-z & a \\ -a & 2-z \end{vmatrix} = (1-z)(2-z) + a^2 = z^2 - 3z + 2 + a^2\,,$$

de discriminant $\Delta = 9 - 4(2 + a^2) = 1 - 4a^2 = (1 - 2a)(1 + 2a)$. Si $|a| > 1/2$, le discriminant est négatif, les racines de C_M ne sont pas réelles, donc f n'est pas diagonalisable. Supposons $|a| \leqslant 1/2$.

Premier cas : $|a| < 1/2$. Le discriminant est positif, le polynôme caractéristique a deux racines réelles distinctes, donc il y a deux valeurs propres distinctes $\lambda = (1/2)(3 + \sqrt{1 - 4a^2})$ et $\mu = (1/2)(3 - \sqrt{1 - 4a^2})$: la transformation f est donc diagonalisable.

Un vecteur propre pour la valeur propre λ est une solution du système d'équations $(M - \lambda I_2)X = 0$. Ce système est de rang 1 car son déterminant est nul : il suffit donc de résoudre une seule équation du système pour trouver X. En posant $X = \begin{bmatrix} x \\ y \end{bmatrix}$, il vient

$$(M - \lambda I_2)X = 0 \iff \begin{bmatrix} 1-\lambda & a \\ -a & 2-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff (1-\lambda)x + ay = 0$$

et le vecteur $U_1 = \begin{bmatrix} a \\ \lambda-1 \end{bmatrix}$ est propre pour la valeur propre λ. De même, le vecteur $U_2 = \begin{bmatrix} a \\ \mu-1 \end{bmatrix}$ est propre pour μ. Dans la base (U_1, U_2), la matrice de f est $\mathrm{diag}(\lambda, \mu)$.

Deuxième cas : $a = 1/2$. Le discriminant Δ est nul et le polynôme

$$C_M = z^2 - 3z + (3/2)^2 = (z - 3/2)^2$$

a une racine double $3/2$. Il n'y a qu'une valeur propre $\lambda = 3/2$, de multiplicité 2. Les vecteurs propres sont les solutions non nulles de l'équation

$$(M-(3/2)I_2)X=0 \Leftrightarrow \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow -x+y=0 \Leftrightarrow \begin{bmatrix} x \\ y \end{bmatrix} = t \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \text{ où } t \in \mathbb{R}.$$

L'ensemble des solutions, c'est-à-dire le sous-espace propre $V(3/2)$, est de dimension 1, engendré par le vecteur propre $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. D'après le critère de diagonalisation, f n'est pas diagonalisable.

Troisième cas : $a = -1/2$. La matrice est la transposée de la précédente, donc le polynôme caractéristique est le même. Le sous-espace propre $V(3/2)$ est encore de dimension 1, engendré par le vecteur propre $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Comme ci-dessus, on en déduit que f n'est pas diagonalisable.

4. Trigonalisation

Comme on vient de le voir, une transformation linéaire n'est pas toujours diagonalisable, même si les racines du polynôme caractéristique appartiennent toutes à $\mathbb{K}$. Dans ce paragraphe, V est un $\mathbb{K}$-espace vectoriel de dimension n.

Un exemple. Soit f une transformation de V. Supposons que, dans une certaine base, la matrice M de f est triangulaire à coefficients diagonaux tous égaux :

$$M = \begin{bmatrix} a & & & \\ 0 & a & & * \\ \vdots & & \ddots & \\ 0 & & \cdots & a \end{bmatrix}, \quad \text{où } a \in \mathbb{K}.$$

Le polynôme caractéristique de f est $(a - z)^n$, la seule racine est a, de multiplicité n, donc a est la seule valeur propre de f.

D'après le critère, f est diagonalisable si et seulement si $\dim V(a) = n$, c'est-à-dire si et seulement si $V(a) = V$. Cette condition signifie que pour tout vecteur $u \in V$, on a $f(u) = au$, autrement dit $f = a\,\mathrm{id}_V$ (ou encore, si $a \neq 0$, f est l'homothétie de rapport a).

Nous avons ainsi montré que f n'est diagonalisable que si c'est la transformation $u \mapsto au$, c'est-à-dire si M est la matrice diagonale aI_n.

Étudions plus précisément le type de transformation linéaire présenté dans cet exemple. Nous continuons à utiliser les notations du paragraphe précédent.

Cas où f a pour seule valeur propre 0

Posons $N_0 = \{\mathbf{0}\}$, $N_1 = \operatorname{Ker} f$, $N_2 = \operatorname{Ker}(f \circ f)$ et en général

$$N_k = \operatorname{Ker}(f^k) \quad \text{pour tout entier } k \geqslant 1,$$

où f^2 est la transformation $f \circ f$ et $f^k = \underbrace{f \circ f \circ \cdots \circ f}_{k \text{ fois}}$.

Les N_k sont des noyaux de transformations linéaires de V, donc des sous-espaces vectoriels de V. Voici des propriétés de ces sous-espaces.

i) On a $\operatorname{Ker} f = V(0) = N_1 \subset N_2 \subset \cdots \subset N_k \subset N_{k+1} \subset \cdots \subset N_n = V$.

ii) Si $k \geqslant 1$, alors pour tout vecteur $u \in N_k$, on a $f(u) \in N_{k-1}$.

Démonstration. Par définition, $N_1 = \operatorname{Ker} f = \operatorname{Ker}(f - 0\,\mathrm{id}_V)$ est le sous-espace propre $V(0)$ pour la valeur propre 0. On a supposé que 0 est la seule valeur propre de f, donc le polynôme caractéristique de f est $C_f = (-1)^n z^n$. D'après le théorème de Cayley-Hamilton (page 152), on en déduit que pour tout vecteur $u \in V$, on a $\mathbf{0} = C_f(f)(u) = (-1)^n f^n(u)$, donc $f^n(u) = \mathbf{0}$. Ainsi tout vecteur de V appartient à N_n, autrement dit $N_n = V$.

Montrons que les sous-espaces N_k sont emboîtés. Soient $k \geqslant 0$ et $u \in N_k$. On a $f^{k+1} = f \circ f^k$ et $f^{k+1}(u) = f\big(f^k(u)\big) = f(\mathbf{0}) = \mathbf{0}$, car $f^k(u) = \mathbf{0}$. Tout vecteur u de N_k appartient donc au noyau de f^{k+1} qui est N_{k+1}. Cela démontre l'inclusion $N_k \subset N_{k+1}$.

Montrons la dernière propriété. Soit $k \geqslant 1$ et $u \in N_k$. On a $f^{k-1} \circ f = f^k$, donc $f^{k-1}\big(f(u)\big) = f^k(u)$; puisque $f^k(u) = \mathbf{0}$, il vient $f^{k-1}\big(f(u)\big) = \mathbf{0}$, autrement dit le vecteur $f(u)$ appartient à N_{k-1}. ■

Construction d'une base adaptée aux N_k

Construisons une base de V dont les premiers vecteurs sont dans $\operatorname{Ker} f = N_1$, les suivants dans N_2, etc, les derniers dans $N_n = V$. Nous dirons qu'une telle base de V est *adaptée* aux sous-espaces emboîtés N_k.

Pour cela, on choisit une base de $V(0) = \operatorname{Ker} f$; si l'on a $\dim(\operatorname{Ker} f) < \dim N_2$, on complète en une base de N_2, grâce au théorème de la base incomplète ; ensuite, si $\dim N_2 < \dim N_3$, on complète en une base de N_3, et l'on continue ainsi jusqu'à obtenir une base de $N_n = V$.

Soit $(u_1, u_2, \ldots, u_n)$ une base de V adaptée aux noyaux N_k. Si le sous-espace propre $V(0)$ est de dimension d, alors $u_1, \ldots, u_d$ forment une base de $V(0)$ et $f(u_i) = 0$ pour $1 \leqslant i \leqslant d$.

Soit u_p un vecteur de la base tel que $p > d$. Il y a un plus petit indice k tel que $u_p \in N_k$. Par construction d'une base adaptée, les vecteurs de N_{k-1} sont combinaison linéaire de vecteurs u_i tels que $i < p$. Or nous savons (propriété (ii)) que $f(u_p)$ appartient à N_{k-1}, donc $f(u_p)$ est combinaison linéaire de $u_1, \ldots, u_{p-1}$.

Il s'ensuit que la matrice T de f dans la base $(u_1, u_2, \ldots, u_n)$ est triangulaire supérieure, de la forme

$$T = \begin{bmatrix} 0 & \cdots & 0 & & \\ \vdots & \ddots & \vdots & * & \\ 0 & \cdots & 0 & & \\ \vdots & & \vdots & \ddots & \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}$$

Les d premières colonnes sont nulles et les coefficients diagonaux sont tous égaux à 0. Puisqu'on a $\mathrm{Ker}(f^n) = N_n = V$, la transformation f^n est nulle, donc $T^n = 0$.

Cas où f a une seule valeur propre λ

Posons $g = f - \lambda \mathrm{id}_V$. Si f a pour matrice M dans une base de V, alors pour tout scalaire $t \in \mathbb{K}$, la transformation $g - t\,\mathrm{id}_V = f - (\lambda + t)\,\mathrm{id}_V$ a pour matrice $M - (\lambda + t)I_n$. Puisque λ est la seule valeur propre de f, cette matrice est inversible sauf pour $t = 0$: la seule valeur propre de g est donc 0. D'après l'étude précédente, il existe une base de V dans laquelle g a pour matrice T. Comme on a $f = g + \lambda\,\mathrm{id}_V$, la matrice de f dans cette base est $T + \lambda I_n$, de la forme

$$T(\lambda) = \begin{bmatrix} \boxed{\lambda I_d} & & * & \\ & \lambda & * & \\ 0 & & \ddots & \\ & & & \lambda \end{bmatrix}$$

Cette matrice est triangulaire supérieure à coefficients diagonaux tous égaux à λ.

Cas général

Supposons que le polynôme caractéristique de f a toutes ses racines dans $\mathbb{K}$ et décomposons-le en produit de facteurs

$$C_f = (\lambda_1 - z)^{m(\lambda_1)}(\lambda_2 - z)^{m(\lambda_2)} \cdots (\lambda_k - z)^{m(\lambda_k)}$$

où les λ_i sont les différentes valeurs propres.

On démontre que si $\mathscr{B}_i$ est une base de $\mathrm{Ker}\left[(f - \lambda_i\,\mathrm{id}_V)^{m(\lambda_i)}\right]$, alors la réunion des $\mathscr{B}_i$ est une base de V. La matrice de f dans cette base est constituée de k matrices carrées de taille $m(\lambda_1), \ldots, m(\lambda_k)$ placées en diagonale. En choisissant pour $\mathscr{B}_i$ une base adaptée aux noyaux

$$\mathrm{Ker}(f - \lambda_i\,\mathrm{id}_V) \subset \mathrm{Ker}\left[(f - \lambda_i\,\mathrm{id}_V)^2\right] \subset \cdots \subset \mathrm{Ker}\left[(f - \lambda_i\,\mathrm{id}_V)^{m(\lambda_i)}\right],$$

ces matrices seront de la forme $T(\lambda_i)$. Énonçons ce résultat que nous admettons.

Proposition. *Si le polynôme caractéristique de f a toutes ses racines $\lambda_1, \ldots, \lambda_k$ dans $\mathbb{K}$, il existe une base de V dans laquelle la matrice de f a une forme « triangulaire par blocs »*

$$\operatorname{diag}(T_1, T_2, \ldots, T_k) = \begin{bmatrix} \boxed{T_1} & & & \\ & \boxed{T_2} & 0 & \\ & 0 & \ddots & \\ & & & \boxed{T_k} \end{bmatrix}$$

où $T_i = T(\lambda_i)$ est une matrice carrée triangulaire supérieure de taille $m(\lambda_i)$ dont les coefficients diagonaux sont tous égaux à λ_i.

La matrice ci-dessus est triangulaire. En conséquence, la proposition affirme que si les racines du polynôme caractéristique de f sont toutes dans $\mathbb{K}$, il y a une base dans laquelle la matrice de f est triangulaire, les coefficients diagonaux étant constitués des valeurs propres répétées autant de fois que leur multiplicité.

Remarque

On peut toujours trigonaliser une matrice à coefficients réels avec des valeurs propres et des vecteurs propres complexes.

5. Applications

5.1 Calcul des puissances d'une matrice

Pour étudier les itérés d'une transformation linéaire, on a souvent besoin de calculer explicitement la puissance p-ième d'une matrice carrée. Voici les principaux exemples où l'on peut effectuer ce calcul de façon assez simple.

La matrice est diagonale

Si $D = \operatorname{diag}(t_1, t_2, \ldots, t_n)$, alors $D^p = \operatorname{diag}(t_1^p, t_2^p, \ldots, t_n^p)$ pour tout entier $p \geqslant 1$.

La matrice est triangulaire et les coefficients diagonaux sont tous nuls

Si la matrice est triangulaire supérieure de taille n, elle est de la forme $T = \begin{bmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{bmatrix}$.

On a $T\boldsymbol{E_1} = 0$ et pour tout $i = 2, \ldots, n$, la i-ème colonne $T\boldsymbol{E_i}$ est combinaison linéaire de $\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots, \boldsymbol{E_{i-1}}$. Ainsi, $T\boldsymbol{E_2}$ est colinéaire à $\boldsymbol{E_1}$ et plus généralement $T^{i-1}\boldsymbol{E_i}$ est colinéaire à $\boldsymbol{E_1}$, donc $T^i\boldsymbol{E_i} = 0$. *A fortiori*, on a $T^n\boldsymbol{E_i} = 0$ et comme cela est vrai quel que soit i, il vient $T^n = 0$. Par conséquent, on a $T^p = 0$ si $p \geqslant n$, et les seules puissances à calculer sont $T^2, \ldots, T^{n-1}$.

Une matrice carrée non nulle dont une puissance est nulle s'appelle une matrice *nilpotente*. Une matrice triangulaire n'ayant que des zéros sur la diagonale est nilpotente.

La matrice est triangulaire à coefficients diagonaux tous égaux

Il s'agit par exemple d'une matrice $T = \begin{bmatrix} \lambda & & * \\ & \ddots & \\ 0 & & \lambda \end{bmatrix}$. On a $T = \lambda I_n + N$, où N est triangulaire supérieure à coefficients diagonaux tous nuls. D'après ce qu'on vient de voir, on a $N^p = 0$ si $p \geqslant n$. Pour calculer T^p, utilisons la formule du binôme de Newton, ce qui est licite car la matrice λI_n commute avec n'importe quelle matrice de $\mathscr{M}_n(\mathbb{K})$:

$$(*) \qquad T^p = (\lambda I_n + N)^p = \lambda^p I_n + \binom{p}{1}\lambda^{p-1}N + \cdots + \binom{p}{k}\lambda^{p-k}N^k + \cdots + N^p$$

Dans cette égalité, les seuls termes éventuellement non nuls sont relatifs aux puissances de N inférieures à n : quel que soit p, il y a donc au plus n termes non nuls.

La matrice est triangulaire par blocs

La forme est $T = \operatorname{diag}(T_1, T_2, \ldots, T_k)$, où les matrices T_i sont triangulaires et occupent des blocs de lignes et de colonnes disjoints les uns des autres, comme dans la proposition page 182. D'après la règle du produit matriciel, on a alors

$$T^p = \operatorname{diag}(T_1^p, T_2^p, \ldots, T_k^p) \quad \text{pour tout entier } p \geqslant 1.$$

Méthode de calcul des puissances

On veut calculer les puissances d'une matrice A. Supposons qu'on dispose d'une matrice inversible P et qu'on sache calculer les puissances de la matrice $B = P^{-1}AP$. En multipliant cette égalité à gauche par P et à droite par P^{-1}, il vient $A = PBP^{-1}$. Par suite, $A^2 = PBP^{-1}PBP^{-1} = PB^2P^{-1}$ et plus généralement $A^p = PB^pP^{-1}$ pour tout entier $p \geqslant 1$. Connaissant B^p, on en déduit A^p en calculant un produit de trois matrices.

Pour calculer les puissances de A, on peut chercher une base $\mathscr{B}$ de $\mathbb{K}^n$ dans laquelle la matrice de la transformation $X \mapsto AX$ est diagonale ou à défaut triangulaire par blocs.

En prenant pour P la matrice de passage de la base canonique à la base $\mathscr{B}$, la matrice de $X \mapsto AX$ dans la base $\mathscr{B}$ est $B = P^{-1}AP$, d'après la formule de changement de base. On aura donc $A^p = PB^pP^{-1}$ pour tout entier $p \geqslant 1$.

Exemple : suite vérifiant une relation de récurrence linéaire. Soient a et b des nombres de $\mathbb{K}$ et soit (u_p) une suite telle que $u_{p+2} = au_{p+1} + bu_p$ pour tout entier $p \geqslant 0$.
Posons $v_p = u_{p+1}$. On a $v_{p+1} = u_{p+2} = av_p + bu_p$, donc

$$\begin{bmatrix} u_{p+1} \\ v_{p+1} \end{bmatrix} = \begin{bmatrix} v_p \\ bu_p + av_p \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix} \begin{bmatrix} u_p \\ v_p \end{bmatrix}$$

ou encore

$$X_{p+1} = AX_p \ , \quad \text{avec } X_p = \begin{bmatrix} u_p \\ v_p \end{bmatrix} \text{ et } A = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix}.$$

On en déduit $X_p = A^pX_0$. Pour calculer les nombres u_p, il suffit donc d'expliciter la matrice A^p et de se donner les valeurs initiales u_0 et $u_1 = v_0$.

Faisons ce calcul pour $a = 6$ et $b = -9$.

Trigonalisation de A. Le polynôme caractéristique de A est $\begin{vmatrix} -z & 1 \\ -9 & 6-z \end{vmatrix} = z^2 - 6z + 9 =$

$(z-3)^2$. La seule valeur propre est 3, racine double du polynôme caractéristique. Puisque A n'est pas la matrice $3I_2$, on en déduit que A n'est pas diagonalisable (exemple page 179). Pour trouver un vecteur propre, on résout le système linéaire :

$$(A - 3I_2)X = 0 \iff \begin{cases} -3x + y = 0 \\ -9x + 3y = 0 \end{cases} \iff y = 3x$$

donc $U_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ est vecteur propre. Complétons ce vecteur en une base de $\mathbb{R}^2$ en posant $U_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. On a $(A - 3I_2)U_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} = U_1$, donc $AU_2 = U_1 + 3U_2$. Puisque $AU_1 = 3U_1$, la matrice de $X \mapsto AX$ dans la base (U_1, U_2) est $T = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$. La matrice de passage de la base canonique à la base (U_1, U_2) a pour colonnes U_1, U_2, c'est donc la matrice $P = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}$, d'inverse $P^{-1} = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}$. D'après la formule du changement de base, on a l'égalité matricielle $T = P^{-1}AP$, ou encore $A = PTP^{-1}$.

Calcul des puissances de A. On a $T = 3I_2 + N$, où la matrice $N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ vérifie $N^2 = 0$. Pour tout entier $p \geqslant 1$, il vient donc

$$T^p = (3I_2 + N)^p = 3^p I_2 + p3^{p-1}N = \begin{bmatrix} 3^p & p3^{p-1} \\ 0 & 3^p \end{bmatrix}$$

Puisque $A^p = PT^pP^{-1}$, on obtient finalement

$$A^p = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 3^p & p3^{p-1} \\ 0 & 3^p \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} (1-p)3^p & p3^{p-1} \\ -p3^{p+1} & (1+p)3^p \end{bmatrix}.$$

Étude de la suite (u_p). Revenons à la suite (u_p) : puisque $\begin{bmatrix} u_p \\ v_p \end{bmatrix} = A^p \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}$, il vient $u_p = (1-p)3^p u_0 + p3^{p-1}u_1$ pour tout entier $p \geqslant 1$.

- Le quotient $\dfrac{u_p}{p3^{p-1}} = u_1 + 3\dfrac{1-p}{p}u_0$ a pour limite $u_1 - 3u_0$: on en déduit que si $u_1 \neq 3u_0$, alors u_p est équivalent à $(u_1 - 3u_0)p3^{p-1}$ quand p tend vers l'infini.
- Si $u_1 = 3u_0$, le vecteur initial $X_0 = \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}$ est colinéaire au vecteur propre U_1, donc $A^pX_0 = 3^pX_0$ et $u_p = 3^pu_0$ (ce qui se vérifie aussi sur l'expression générale de u_p).

Nous présenterons dans le paragraphe suivant d'autres exemples de calculs d'une puissance de matrice.

5.2 Étude d'itérations linéaires

Soit A une matrice carrée de taille n. Étant donné un vecteur $X_0 \in \mathbb{K}^n$, les itérés de X_0 par la transformation $X \mapsto AX$ sont les vecteurs $X_1 = AX_0, X_2 = AX_1 = A^2X_0, \ldots, X_p = AX_{p-1} = A^pX_0, \ldots$

Dans le cas particulier où X_0 est un vecteur propre pour la valeur propre λ, on a simplement $X_1 = \lambda X_0, X_2 = \lambda A X_0 = \lambda^2 X_0, \ldots, X_p = \lambda^p X_0$. Si $|\lambda| < 1$, alors λ^p tend vers 0 quand p tend vers l'infini et les X_p tendent vers le vecteur nul.
Voici un résultat général.

Proposition. *Soit $A \in \mathscr{M}_n(\mathbb{K})$. Si toutes les racines du polynôme caractéristique de A sont de module strictement inférieur à 1, alors A^p tend vers 0 quand p tend vers l'infini.*

Démonstration. Supposons d'abord que $A = \lambda I + N$, où N est triangulaire avec coefficients diagonaux tous nuls. Le nombre λ est donc valeur propre de A. Puisque $N^n = 0$, la formule $(*)$ page 183 s'écrit pour tout $p \geqslant n$:

$$A^p = \lambda^p I_n + \binom{p}{1}\lambda^{p-1}N + \cdots + \binom{p}{k}\lambda^{p-k}N^k + \cdots + \binom{p}{n-1}\lambda^{p-n+1}N^{n-1}.$$

Si $\lambda = 0$, alors $A^p = 0$ pour $p \geqslant n$, donc A^p tend vers 0. Supposons $\lambda \neq 0$. On a $\binom{p}{k} = \frac{1}{k!}p(p-1)\cdots(p-k+1) \leqslant p^k$, donc $\left|\binom{p}{k}\lambda^{p-k}\right| \leqslant |\lambda|^{-k}p^k|\lambda|^p$. Quand p tend vers l'infini, on sait que, pour k fixé, $p^k|\lambda|^p$ tend vers 0, car on a supposé $|\lambda| < 1$: les coefficients de la matrice $\binom{p}{k}\lambda^{p-k}N^k$ ont donc pour limite 0 quand p tend vers l'infini. Pour tout $p \geqslant n$, A^p est une somme de n matrices qui tendent vers 0, donc A^p tend vers 0 quand p tend vers l'infini.
Supposons maintenant que A est constituée de blocs diagonaux $\lambda_i + N_i$ de la forme qu'on vient de traiter. Les λ_i sont des valeurs propres et A^p s'obtient en élevant chaque bloc à la puissance p. Puisque dans A^p, chaque bloc diagonal tend vers 0 quand p tend vers l'infini, il en va de même de A^p.
Dans le cas général, trigonalisons A sur $\mathbb{C}$: on obtient une matrice $P \in \mathscr{M}_n(\mathbb{C})$ et une matrice T triangulaire par blocs telles que $A^p = PT^pP^{-1}$ quel que soit l'entier $p \geqslant 1$. On vient de montrer que T^p tend vers 0 quand p tend vers l'infini. Chaque coefficient du produit PT^pP^{-1} est une combinaison linéaire, à coefficients indépendants de p, des coefficients de T^p : les coefficients de A^p tendent donc vers 0 quand p tend vers l'infini. ■

Corollaire. *Soient $A \in \mathscr{M}_n(\mathbb{K})$ et B un vecteur-colonne de $\mathbb{K}^n$.*

- *Si 1 n'est pas racine du polynôme caractéristique de A, la transformation $X \mapsto AX + B$ a un unique point fixe $W = (I_n - A)^{-1}B$.*
- *Supposons que le polynôme caractéristique de A a toutes ses racines de module strictement inférieur à 1. Alors toute suite (X_p) telle que $X_{p+1} = AX_p + B$ a pour limite le point fixe W tel que $W = AW + B$.*

Démonstration. Un vecteur $X \in \mathbb{K}^n$ est point fixe si et seulement si $AX + B = X$, ce qui s'écrit $(I_n - A)X = B$. Si 1 n'est pas racine du polynôme caractéristique de A, la matrice $I_n - A$ est inversible, donc l'équation a pour seule solution $X = (I_n - A)^{-1}B$. Supposons que le polynôme caractéristique de A a toutes ses racines de module strictement inférieur à 1. En particulier, 1 n'est pas racine, donc il existe un unique point fixe W, tel que $AW + B = W$. Soit (X_p) une suite telle que $X_{p+1} = AX_p + B$. On a $X_{p+1} - W = (AX_p + B) - (AW + B) = A(X_p - W)$, donc $X_p - W = A^p(X_0 - W)$ pour tout entier $p \geqslant 1$. Quand p tend vers l'infini, A^p tend vers 0 d'après la proposition précédente, donc X_p tend vers W. ■

Itération dans $\mathbb{R}^2$

Soit $A \in \mathscr{M}_2(\mathbb{R})$ et soit $f : \mathbb{R}^2 \to \mathbb{R}^2$ la transformation $X \mapsto AX$.

Exemple 1. Supposons que le polynôme caractéristique de A a deux racines complexes distinctes. Ce polynôme étant à coefficients réels, les deux racines sont conjuguées : notons-les $\lambda = a + b\boldsymbol{i}$ et $\bar{\lambda} = a - b\boldsymbol{i}$, où $b \neq 0$.

Soit $U \in \mathbb{C}^2$ un vecteur propre pour la valeur propre λ. Puisque A est à coefficients réels, on a $A\overline{U} = \overline{AU} = \overline{\lambda U} = \bar{\lambda}\overline{U}$, donc $\overline{U}$ est vecteur propre pour $\bar{\lambda}$.

Posons $U = U_1 + \boldsymbol{i}\,U_2$, où U_1 et U_2 sont des vecteurs à coefficients réels.

Les vecteurs U_1 et U_2 forment une base de $\mathbb{R}^2$.

> On a en effet $\overline{U} = U_1 - \boldsymbol{i}U_2$, $2U_1 = U + \overline{U}$ et $2\boldsymbol{i}U_2 = U - \overline{U}$; si x et y sont des nombres réels tels que $xU_1 + yU_2 = 0$, alors $0 = 2xU_1 + 2yU_2 = x(U + \overline{U}) - y\boldsymbol{i}(U - \overline{U}) = (x - y\boldsymbol{i})U + (x + y\boldsymbol{i})\overline{U}$. Puisque $\lambda \neq \bar{\lambda}$, les vecteurs propres U et $\overline{U}$ sont indépendants dans $\mathbb{C}^2$, donc $x + y\boldsymbol{i} = 0$ et $x = y = 0$.

Pour étudier les itérés d'un vecteur par f, utilisons les coordonnées dans la base (U_1, U_2). On a $AU_1 + \boldsymbol{i}\,AU_2 = AU = \lambda U = (a + b\boldsymbol{i})(U_1 + \boldsymbol{i}U_2) = aU_1 - bU_2 + \boldsymbol{i}(bU_1 + aU_2)$, d'où en séparant parties réelle et imaginaire $AU_1 = aU_1 - bU_2$ et $AU_2 = bU_1 + aU_2$. La matrice de f dans la base (U_1, U_2) est donc $B = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$.

Il faut maintenant calculer B^p. Posons $r = \sqrt{a^2 + b^2}$; puisque $b \neq 0$, on a $r > 0$ et il existe un nombre $\theta \in [0, 2\pi[$ tel que $a = r\cos\theta$ et $b = r\sin\theta$. Il vient $B = rM(\theta)$, avec $M(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$, et par suite (exemple page 170)

$$B^p = r^p \begin{bmatrix} \cos p\theta & \sin p\theta \\ -\sin p\theta & \cos p\theta \end{bmatrix}.$$

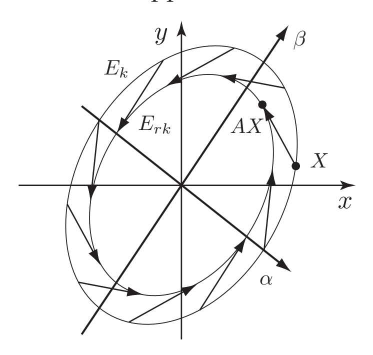
Notons (α, β) les coordonnées d'un vecteur quelconque dans la base (U_1, U_2) : on a ainsi $\begin{bmatrix} x \\ y \end{bmatrix} = P \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, où la matrice de passage P a pour colonnes U_1, U_2. Donnons-nous un vecteur initial $X_0 = \alpha_0 U_1 + \beta_0 U_2$ et notons $X_p = \alpha_p U_1 + \beta_p U_2$ les itérés de X_0 par la transformation $X \mapsto AX$. Pour tout entier $p \geqslant 1$, on a $\begin{bmatrix} \alpha_p \\ \beta_p \end{bmatrix} = B^p \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}$.

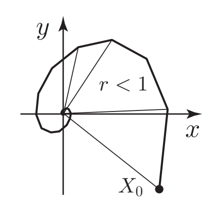
Si l'on pose $\begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} = B \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} r\alpha\cos\theta + r\beta\sin\theta \\ -r\alpha\sin\theta + r\alpha\cos\theta \end{bmatrix}$, alors on a

$$(*) \qquad \alpha'^2 + \beta'^2 = r^2(\alpha^2 + \beta^2)$$

Pour tout nombre réel $k > 0$, la courbe E_k formée des points $(x, y) \in \mathbb{R}^2$ dont les coordonnées α, β vérifient $\alpha^2 + \beta^2 = k^2$ est une ellipse centrée à l'origine. D'après $(*)$, tout point $X \in E_k$ est transformé en un point $AX \in E_{rk}$.

Sur la figure ci-dessous, nous avons représenté les axes de coordonnées α et β, dirigés par U_1 et U_2, et nous avons supposé $r < 1$.





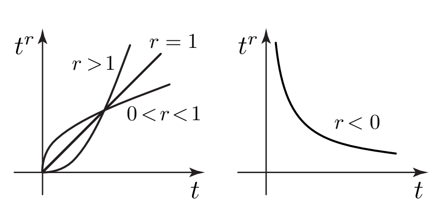
Les itérés de X_0 sont sur une spirale qui tend vers l'origine si $r<1$, qui s'en écarte si $r>1$. Noter que l'angle des vecteurs X, AX n'est pas θ en général.
Nous avons déjà rencontré une trajectoire de cette forme page 17, dans un exemple d'itération.

Exemple 2. Supposons que le polynôme caractéristique de A a deux racines réelles λ, μ.
Il y a une base (U, V) de vecteurs propres associés et tout point $X \in \mathbb{R}^2$ a des coordonnées α, β dans cette base : si $X = \alpha U + \beta V$, alors AX a pour coordonnées $\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}=\begin{bmatrix} \lambda\alpha \\ \mu\beta \end{bmatrix}$ dans la base (U,V) et les itérés d'un point initial $X_0=\alpha_0 U+\beta_0 V$ ont pour coordonnées $\begin{bmatrix} \alpha_p \\ \beta_p \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}^p \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \lambda^p \alpha_0 \\ \mu^p \beta_0 \end{bmatrix}$. Pour trouver une équation de la courbe où se trouvent ces points, éliminons p entre les égalités $\begin{cases} \alpha_p = \lambda^p \alpha_0 \\ \beta_p = \mu^p \beta_0 \end{cases}$. Pour simplifier, supposons λ et μ strictement positifs et plaçons nous dans le cas général $\alpha_0 \neq 0$, $\beta_0 \neq 0$, $\lambda \neq 1$ et $\mu \neq 1$. On peut tirer p de la première égalité et porter dans la seconde :

$$\lambda^p = \frac{\alpha_p}{\alpha_0} \quad , \quad p = (\ln\lambda)^{-1} \ln\left(\frac{\alpha_p}{\alpha_0}\right) \quad , \quad \ln\left(\frac{\beta_p}{\beta_0}\right) = p \ln\mu = \frac{\ln\mu}{\ln\lambda} \ln\left(\frac{\alpha_p}{\alpha_0}\right)$$

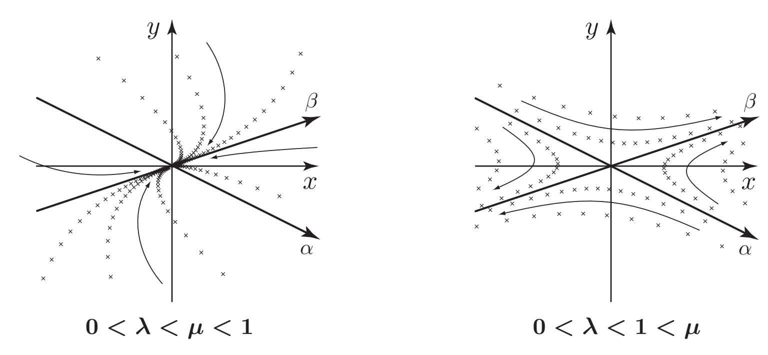
$$\frac{\beta_p}{\beta_0} = \left(\frac{\alpha_p}{\alpha_0}\right)^r \quad , \text{ où l'on a posé } r = \frac{\ln\mu}{\ln\lambda}.$$

Pour des valeurs α_0 et β_0 positives, les itérés se situent sur la courbe d'équation $\beta = k\alpha^r$, où $k = \beta_0/\alpha_0^r$; l'exposant r ne dépend que des valeurs propres. Les figures ci-contre montrent l'allure de la fonction puissance $t \mapsto t^r$ selon les valeurs du nombre r.



Si l'on change le signe de α_0, alors α_p change de signe, de même pour β_0 et β_p : cela fournit des courbes symétriques par rapport aux axes. Les figures ci-après montrent le comportement des itérés selon les valeurs propres de A : les droites correspondent

aux directions propres et les flèches indiquent le sens de parcours.



$\mathbf{0 < \lambda < \mu < 1}$ $\mathbf{0 < \lambda < 1 < \mu}$

Dans le cas $0 < \lambda < \mu < 1$, les itérés tendent bien vers l'origine. On peut voir ces mêmes trajectoires d'itération page 19.

5.3 Suite de transitions probabilistes

Considérons un phénomène aléatoire pouvant prendre un nombre fini d'états, les changements d'états se faisant successivement. Numérotons les états de 1 à k, appelons X_0 le numéro de l'état initial et X_n le numéro de l'état après n changements. À la variable aléatoire X_n, associons le vecteur

$$Z(n) = \begin{bmatrix} p(X_n = 1) \\ p(X_n = 2) \\ \vdots \\ p(X_n = k) \end{bmatrix},$$

où $p(X_n\!=\!i)$ est la probabilité pour qu'on soit dans l'état i après n changements d'état.

Faisons l'hypothèse suivante : la probabilité pour passer de l'état j à l'état i reste constante, c'est-à-dire ne dépend pas de n.

Pour tous entiers i, j compris entre 1 et k, notons p_{ij} la probabilité de transiter de l'état j à l'état i. On a donc $p_{ij} = p(X_{n+1} = i \mid X_n = j)$, probabilité conditionnelle pour que $X_{n+1} = i$ sachant que $X_n = j$.
On définit la *matrice de transitions* en posant $M = [p_{ij}]$, où comme d'habitude le second indice est celui de la colonne.
Par définition d'une probabilité conditionnelle, on a

$$p(X_{n+1}\!=\!i)\!=\!p_{i1}p(X_n\!=\!1)\!+\!p_{i2}p(X_n\!=\!2)\!+\!\cdots\!+\!p_{ik}p(X_n\!=\!k)\!=\![\,p_{i1}\ \ p_{i2}\ \ \cdots\ \ p_{ik}\,]Z(n),$$

autrement dit pour tout entier $n \geqslant 1$, on a $Z(n+1) = MZ(n)$.

Si l'état initial porte le numéro $X_0 = j$, alors $Z(n) = M^n \boldsymbol{E_j}$,
où $\boldsymbol{E_j}$ est le j-ème vecteur canonique de $\mathbb{R}^k$.

Exemple. Dans une population exposée à une maladie transmissible non mortelle, distinguons trois catégories : les malades, les individus qui sont infectés mais ne développent pas la maladie et ceux qui restent sains. Une étude statistique montre que

- sur 100 individus sains, 30 tombent malades la semaine suivante, 20 deviennent simplement infectés et 50 restent sains ;
- sur 100 individus infectés, 70 développent la maladie dans la semaine ; de plus, un individu infecté ne redevient pas sain immédiatement, faute de soins ;
- sur 100 malades, 40 restent infectés la semaine suivante et 35 sont guéris.

Formons la matrice de transitions (les changements d'état étant hebdomadaires) : la première colonne correspond aux transitions depuis l'état « malade » (M), la deuxième depuis l'état « infecté » (I) et la troisième depuis l'état « sain » (S).
Les coefficients de la première colonne sont donc : la probabilité $p_{11} = 0{,}25$ de rester à l'état (M), la probabilité $p_{21} = 0{,}4$ de passer de l'état (M) à l'état (I) et la probabilité $p_{31} = 0{,}35$ de passer (M) à (S).
De même, nous écrivons dans la deuxième colonne les probabilités p_{12}, p_{22} et p_{32} de passer de l'état (I) aux états (M), (I), (S), et dans la troisième colonne les probabilités de transition de (S) vers (M), (I), (S). On obtient $M = \begin{bmatrix} 0{,}25 & 0{,}70 & 0{,}30 \\ 0{,}40 & 0{,}30 & 0{,}20 \\ 0{,}35 & 0 & 0{,}50 \end{bmatrix}$.
Dans chaque colonne, la somme des coefficients vaut évidemment 1.

Propriétés d'une matrice de transitions

Soit $M = [p_{ij}]$ une matrice de transitions de taille k.

Propriété 1. Les coefficients de M sont positifs ou nuls et dans chaque colonne, la somme des coefficients vaut 1.

En effet, les coefficients de la j-ième colonne sont les différentes probabilités de transition depuis l'état j : leur somme est donc 1.

Cela veut dire que si ${}^tU$ est le vecteur-ligne dont tous les coefficients sont égaux à 1, on a $({}^tU)M = {}^tU$, ou encore en transposant $({}^tM)U = U$. Ainsi 1 est valeur propre de ${}^tM$, donc de M, car des matrices transposées ont même polynôme caractéristique (page 152). Énonçons cette propriété.

Propriété 2. La transformation $X \mapsto MX$ a pour valeur propre 1.

Propriété 3. Les valeurs propres de M sont de valeur absolue inférieure ou égale à 1.

Soit λ une valeur propre de M. Puisque ${}^tM$ et M ont mêmes valeurs propres, λ est valeur propre de ${}^tM$, donc il existe un vecteur-colonne non nul V tel que $({}^tM)V = \lambda V$, ou encore, en transposant, $({}^tV)M = \lambda {}^tV$. Notons $v_1, v_2, \ldots, v_k$ les coordonnées de V et soit v_q une coordonnée de plus grande valeur absolue, donc $|v_j| \leqslant |v_q|$ quel que soit $j = 1, 2, \ldots, k$. Posons $[\, w_1 \ w_2 \ \cdots \ w_k \,] = ({}^tV)M$. On a pour tout j

$$\begin{aligned} |w_j| = |v_1 p_{1j} + v_2 p_{2j} + \cdots + v_k p_{kj}| &\leqslant |v_1| p_{1j} + |v_2| p_{2j} + \cdots + |v_k| p_{kj} \\ &\leqslant |v_q| p_{1j} + |v_q| p_{2j} + \cdots + |v_q| p_{kj} = |v_q| \,. \end{aligned}$$

Puisqu'on a par hypothèse $w_j = \lambda v_j$ pour tout j, il vient $|\lambda||v_q| = |\lambda v_q| = |w_q| \leqslant |v_q|$. Or v_q n'est pas nul, car V n'est pas le vecteur nul. Donc $|\lambda| \leqslant 1$, ce qui démontre la propriété (3).

Étude d'une suite de transitions

Supposons pour simplifier que 1 est valeur propre simple de M, que -1 n'est pas valeur propre et que M est diagonalisable. Notons V un vecteur propre de M pour la valeur propre 1.

D'après ces hypothèses, la matrice de $X \mapsto MX$ dans une base convenable de vecteurs propres est de la forme $D = \text{diag}(1, \lambda_2, \ldots, \lambda_k)$, avec $|\lambda_i| < 1$ pour tout $i \geqslant 2$. En appelant Q la matrice des vecteurs propres, on a $M = QDQ^{-1}$ et donc $M^n = QD^nQ^{-1}$. Quand n tend vers l'infini, λ_i^n tend vers 0, donc D^n tend vers $\Delta = \text{diag}(1, 0, \ldots, 0)$ et QD^n tend vers $Q\Delta$. La première colonne de $Q\Delta$ est la première colonne de Q, c'est-à-dire le vecteur propre V, et les autres colonnes sont nulles. Il s'ensuit que $M^n = QD^nQ^{-1}$ tend vers la matrice $Q\Delta Q^{-1} = [\, V \; 0 \; \cdots \; 0 \,] \, Q^{-1}$ dont toutes les colonnes sont colinéaires à V. Puisque M^n est une matrice de transitions, la somme des coefficients d'une colonne quelconque vaut 1. On en déduit :

la limite de M^n est la matrice dont toutes les colonnes sont égales à $(1/s)V$, où s la somme des coefficients de V.

Reprenons la matrice de transitions de l'exemple précédent.
Le polynôme caractéristique de M est $(1/200)(z-1)(200z^2 - 10z - 17)$ et les valeurs propres sont 1, $(1/40)(1 \pm \sqrt{137})$, ces deux dernières valeurs valant à peu près 0,317 et $-0{,}267$. À partir de $n = 5$, les deux premières décimales des coefficients de M^n restent stables : à 10^{-2} près, on obtient

$$\lim_{n \mapsto +\infty} (M^n) = \begin{bmatrix} 0{,}40 & 0{,}40 & 0{,}40 \\ 0{,}31 & 0{,}31 & 0{,}31 \\ 0{,}28 & 0{,}28 & 0{,}28 \end{bmatrix}$$

Qu'un individu soit à l'origine malade, infecté ou sain, la probabilité pour qu'au bout de six semaines il soit encore malade est d'environ 0,4, la probabilité pour qu'il soit seulement infecté est d'environ 0,31 et la probabilité pour qu'il soit sain est d'un peu plus de 0,28. La colonne V de la matrice $\lim_{n \mapsto +\infty} M^n$ est vecteur propre de M pour la valeur propre 1 : on a $MV = V$. Cela veut dire que la répartition 40% de malades, 31% d'infectés et 29% de sains reste stable au cours du temps. Ce que l'on a montré, c'est que quelles que soient les proportions initiales de sujets malades, infectés ou sains, la population évolue probablement vers cette répartition stable.

5.4 Itérations affines commandables

On rencontre souvent des transformations de $\mathbb{R}^n$ de la forme $X \mapsto AX + V$, où la matrice $A \in \mathscr{M}_n(\mathbb{R})$ est constante et où V est un vecteur de $\mathbb{R}^n$ dont on peut faire varier les coordonnées.

Supposons que les coordonnées de V dépendent linéairement de m paramètres réels $u_1, u_2, \ldots, u_m$: nous dirons que le vecteur $U = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}$ est le *vecteur de contrôle*. Notre hypothèse est que V dépend du vecteur de contrôle par une relation de la forme $V = BU$, où B est une matrice à n lignes et m colonnes. La transformation s'écrit

$$X \mapsto AX + BU$$

et s'appelle un *système linéaire contrôlé*. Si l'on choisit successivement $U_0, U_1, U_2, \ldots, U_{k-1}$ comme vecteurs de contrôle, un vecteur initial X_0 est transformé en

$$X_1 = AX_0 + BU_0 \ , \quad X_2 = AX_1 + BU_1 \ , \quad \ldots \quad , \ X_k = AX_{k-1} + BU_{k-1} \, .$$

Voici l'une des questions qui se pose à propos de la suite des vecteurs X_k :

> à partir d'un vecteur initial quelconque, peut-on atteindre un vecteur objectif donné à l'avance en choisissant convenablement les commandes ?

Définition
Le système linéaire contrôlé $X \mapsto AX + BU$ est *commandable* si pour tous vecteurs X_0 et X_f appartenant à $\mathbb{R}^n$, il existe dans $\mathbb{R}^m$ une suite finie de commandes $U_0, U_1, \ldots, U_{k-1}$ telle que $X_k = X_f$.

Si l'on part du vecteur X_0, on obtient $X_1 = AX_0 + BU_0$, $X_2 = A^2X_0 + ABU_0 + BU_1$ et en général :

$$(*) \qquad X_k = A^kX_0 + A^{k-1}BU_0 + A^{k-2}BU_1 + \cdots + ABU_{k-2} + BU_{k-1}$$

Définition
La *matrice de commandabilité* est la matrice $C = [\,A^{n-1}B \ A^{n-2}B \ \cdots \ AB \ B\,]$ obtenue en juxtaposant les colonnes des n matrices $A^{n-1}B, A^{n-2}B, \ldots, AB, B$. La matrice C possède n lignes et nm colonnes.

Critère de commandabilité. *Le système linéaire contrôlé $X \mapsto AX + BU$ est commandable si et seulement si sa matrice de commandabilité est de rang n.*

Cette propriété signifie que les lignes de la matrice C sont indépendantes, ou que parmi les nm colonnes de C, on peut en trouver n indépendantes.
Pour démontrer le critère, nous avons besoin d'un résultat préliminaire général sur les matrices carrées.

Lemme. *Soit $A \in \mathscr{M}_n(\mathbb{K})$. Pour tout entier $k \geqslant 0$, la matrice A^k est combinaison linéaire des matrices $I_n, A, A^2, \ldots, A^{n-1}$.*

Démonstration. D'après le théorème de Cayley-Hamilton (page 152), il existe des nombres λ_i tels que $A^n + \lambda_1 A^{n-1} + \cdots + \lambda_{n-1}A + \lambda_n I_n = 0$: la matrice A^n est donc combinaison linéaire de $I_n, A, \ldots, A^{n-1}$. Si une matrice A^k est combinaison linéaire de $I_n, A, \ldots, A^{n-1}$,

disons $A^k = \alpha_0 I_n + \alpha_1 A +, \cdots + \alpha_{n-1}A^{n-1}$, alors en multipliant par A, on obtient $A^{k+1} = \alpha_0 A + \alpha_1 A^2 +, \cdots + \alpha_{n-1}A^n$; puisque A^n est combinaison de $I_n, A, \ldots, A^{n-1}$, on en déduit que A^{k+1} est aussi combinaison linéaire de $I_n, A, \ldots, A^{n-1}$. Cela démontre le résultat, d'après le principe de récurrence. ■

Démonstration du critère. Considérons la matrice $C_k = [\,A^{k-1}B \; A^{k-2}B \; \cdots \; AB \; B\,]$ à km colonnes. D'après le lemme, les matrices $A^{k-1}B, \ldots, AB, B$ sont combinaisons linéaires de $A^{n-1}B, \ldots, AB, B$. On en déduit que si $k \geqslant n$, les matrices C_k et C_n ont même rang.
Supposons le système commandable. Tout vecteur $X \in \mathbb{R}^n$ peut être atteint à partir du vecteur nul : pour un certain entier p, on a donc d'après $(*)$

$$X = A^{p-1}BU_0 + A^{p-2}BU_1 + \cdots + ABU_{p-2} + BU_{p-1} = C_p U^*, \text{ où } U^* = \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{p-1} \end{bmatrix}.$$

Le vecteur U^* possède pm coordonnées. En prenant pour X les vecteurs canoniques de $\mathbb{R}^n$ et pour k le plus grand des entiers n et p correspondants, on obtient une matrice C_k dont les colonnes engendrent $\mathbb{R}^n$, donc de rang n. D'après le résultat indiqué en début de démonstration, on en déduit que la matrice $C = C_n$ est aussi de rang n.
Réciproquement, supposons que la matrice de commandabilité C est de rang n et soient X_0, X_f des vecteurs de $\mathbb{R}^n$. Le vecteur $X_f - A^n X_0$ est combinaison linéaire des colonnes de C, donc s'écrit $X_f - A^n X_0 = A^{n-1}BU_0 + A^{n-2}BU_1 + \cdots + ABU_{n-2} + BU_{n-1}$ pour certains vecteurs $U_0, U_1, \ldots, U_{n-1}$ de $\mathbb{R}^m$. Si l'on prend le vecteur X_0 comme vecteur initial, on obtient $X_n = X_f$ d'après $(*)$. Le système est donc commandable. ■

Exemple. Considérons un oscillateur régi par l'équation différentielle

$$(e) \qquad \ddot{x} = -\omega^2 x + u\,,$$

où comme d'habitude, $\dot{x}$ désigne la dérivée par rapport au temps et $\ddot{x}$ la dérivée seconde. Supposons u constant sur un intervalle de temps $[t_0, t_0 + T[$. Alors la fonction constante u/ω^2 est solution. Si t_* est entre t_0 et t_0+T, il y a une unique solution ayant, à cet instant t_*, un état x_* et une vitesse $\dot{x}_*$ données : cette solution est définie par

$$x(t) = (x_* - u/\omega^2)\cos\big[\omega(t-t_*)\big] + (\dot{x}_*/\omega)\sin\big[\omega(t-t_*)\big] + u/\omega^2$$

Discrétisation de l'équation différentielle (e). Donnons-nous x_0 et $\dot{x}_0$ à l'instant $t_0 = 0$ et choisissons une durée $T > 0$. Servons-nous de u comme commande en prenant $u = u_0$ sur l'intervalle de temps $[0, T[$, $u = u_1$ sur $[T, 2T[$, $u = u_{k-1}$ sur $[(k-1)T, kT[$: à l'instant kT, x se trouve alors dans un état x_k avec une vitesse $\dot{x}_k$. En fixant $u = u_k$ pendant l'intervalle de temps $[kT, kT + T[$, la solution pour $kT \leqslant t \leqslant kT + T$ est

$$x(t) = (x_k - u_k/\omega^2)\cos\big[\omega(t-kT)\big] + (\dot{x}_k/\omega)\sin\big[\omega(t-kT)\big] + u_k/\omega^2\,.$$

L'état à l'instant $t = kT+T = (k+1)T$ est donc

$$x_{k+1} = (x_k - u_k/\omega^2)\cos(\omega T) + (\dot{x}_k/\omega)\sin(\omega T) + u_k/\omega^2$$

et la vitesse est donnée par

$$\dot{x}_{k+1} = \dot{x}(kT+T) = -\omega(x_k - u_k/\omega^2)\sin(\omega T) + \dot{x}_k\cos(\omega T)\,.$$

En réunissant état et vitesse en un vecteur de $\mathbb{R}^2$, on obtient les égalités matricielles

$$\begin{bmatrix} x_{k+1} \\ \dot{x}_{k+1} \end{bmatrix} = \begin{bmatrix} \cos\omega T & (1/\omega)\sin\omega T \\ -\omega\sin\omega T & \cos\omega T \end{bmatrix} \begin{bmatrix} x_k - u_k/\omega^2 \\ \dot{x}_k \end{bmatrix} + \begin{bmatrix} u_k/\omega^2 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \cos\omega T & (1/\omega)\sin\omega T \\ -\omega\sin\omega T & \cos\omega T \end{bmatrix} \begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix} + (1/\omega^2) \begin{bmatrix} 1-\cos\omega T \\ \omega\sin\omega T \end{bmatrix} u_k$$

Posons $X_k = \begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix}$, $A = \begin{bmatrix} \cos\omega T & (1/\omega)\sin\omega T \\ -\omega\sin\omega T & \cos\omega T \end{bmatrix}$ et $B = (1/\omega^2)\begin{bmatrix} 1-\cos\omega T \\ \omega\sin\omega T \end{bmatrix}$. Il vient

$$(d) \qquad X_{k+1} = AX_k + Bu_k$$

où $u_0, u_1, \ldots, u_k$ est la suite des commandes.
Le système linéaire à contrôle numérique $X \mapsto AX + Bu$ s'appelle le *discrétisé de l'équation différentielle* (e) *à la période d'échantillonnage* T.
Dans cet exemple, le vecteur de commande est un scalaire et la matrice B n'a qu'une seule colonne. Remarquons que si ωT est un multiple entier de 2π, alors B est nulle et la commande u n'agit pas : il est clair que, dans ce cas, le système n'est pas commandable.

Étude de la commandabilité. La matrice de commandabilité $C = [\,AB \;\; B\,]$ possède deux lignes et deux colonnes. On a

$$AB = (1/\omega^2) \begin{bmatrix} (\cos\omega T)(1-\cos\omega T) + (\sin\omega T)^2 \\ (\cos\omega T)(\omega\sin\omega T) + \omega(\sin\omega T)(\cos\omega T - 1) \end{bmatrix}$$

$$= (\cos\omega T)B + (1/\omega^2)(\sin\omega T) \begin{bmatrix} \sin\omega T \\ \omega(\cos\omega T - 1) \end{bmatrix}$$

et d'après la linéarité du déterminant par rapport aux vecteurs-colonne, il vient

$$\det C = \det(AB, B) = (\cos\omega T)\det(B,B) + (1/\omega^4)(\sin\omega T) \begin{vmatrix} \sin\omega T & 1-\cos\omega T \\ \omega(\cos\omega T - 1) & \omega\sin\omega T \end{vmatrix}$$

$$= (1/\omega^3)(\sin\omega T) \begin{vmatrix} \sin\omega T & 1-\cos\omega T \\ \cos\omega T - 1 & \sin\omega T \end{vmatrix}, \quad \text{car } \det(B,B) = 0.$$

$$\det C = (1/\omega^3)(\sin\omega T)\left[(\sin\omega T)^2 + (1-\cos\omega T)^2\right] = (2/\omega^3)(\sin\omega T)(1-\cos\omega T)$$

Ce déterminant est nul si et seulement si $\omega T = n\pi$, où n est un entier.
D'après le critère de commandabilité, on en déduit que le système linéaire contrôlé (d) est commandable si et seulement si la période d'échantillonnage T n'est pas un multiple entier de la demi-période π/ω de l'oscillateur (e).

Exercices

@ **1. Polynômes prenant des valeurs entières sur les entiers**

a) On considère des polynômes $P_0, P_1, \dots, P_n$ tels que P_i est de degré i. Montrer que ces polynômes sont indépendants.

b) Montrer que tout polynôme P de degré au plus n s'écrit de manière unique $P(z)=a_0+a_1\dfrac{z}{1!}+a_2\dfrac{z(z-1)}{2!}+\cdots+a_n\dfrac{z(z-1)\cdots(z-n+1)}{n!}$, où les a_i sont des nombres.

c) On suppose que dans l'expression ci-dessus, tous les coefficients a_i sont des entiers relatifs. Montrer que si q est un nombre entier quelconque, alors $P(q)$ est entier.

@ **2. Calcul des sommes $1^q + 2^q + \cdots + k^q$.** Notons $\mathbb{P}_n$ l'espace vectoriel des polynômes à coefficients réels et de degré au plus n (donc $\dim \mathbb{P}_n = n+1$).

a) Montrer que si P est un polynôme de degré k, alors le polynôme $P(z)-P(z-1)$ est de degré au plus $k-1$. Dans la suite, on considère l'application $f : \mathbb{P}_n \to \mathbb{P}_n$ qui à tout polynôme $P(z)$ associe le polynôme $P(z)-P(z-1)$.

b) Montrer que l'application f est linéaire.

c) Montrer que le noyau de f est constitué des polynômes constants. En déduire que l'image de f est constituée des polynômes de degré au plus $n-1$.

d) Soit q un entier tel que $0 < q \leqslant n-1$. Montrer qu'il existe un unique polynôme P_q de degré $q+1$ tel que $P_q(z)-P_q(z-1)=z^q$ et $P_q(0)=0$. Calculer les polynômes P_1, P_2, P_3.

e) Montrer que pour tout entier $k \geqslant 1$, on a $1^q + 2^q + \cdots + k^q = P_q(k)$.

@ **3.** Posons $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, où $b \neq 0$.

a) Montrer que les vecteurs propres de A sont les solutions (x, y) non nulles de l'équation $cx^2+(d-a)xy-by^2=0$. En déduire que les pentes des vecteurs propres réels sont solutions de l'équation $bt^2-(d-a)t-c=0$.

b) On suppose que tous les coefficients de la matrice sont des nombres positifs ou nuls.

(i) Montrer que les valeurs propres sont réelles et qu'au moins l'une d'elles est positive ou nulle.

(ii) Montrer qu'il existe un vecteur propre de pente positive ou nulle et à coordonnées positives ou nulles.

Pour une matrice carrée de taille quelconque à coefficients positifs ou nuls, il existe au moins un vecteur propre à coordonnées positives ou nulles, associé à la valeur propre de plus grand module (théorème de Perron-Frobenius).

@ **4.** On considère la matrice $M = \begin{bmatrix} a & b & c & d \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ p & q & r & s \end{bmatrix}$.

a) Trouver deux valeurs propres réelles. Factoriser le polynôme caractéristique en produit de deux polynômes de degré 2.

b) On suppose $a = -s = 3$, $p = 5$ et $d = -1$. Calculer les valeurs propres et les vecteurs propres de M. Montrer que M est diagonalisable. (il y a quatre valeurs propres)

@ **5.** Ces matrices sont diagonalisables : pourquoi ? (les nombres t, a, b, c, x, y, θ sont réels)

$$\begin{bmatrix} e^t & e^t - 1 & e^t - e^{-t} \\ 0 & 1 & 1 - e^{-t} \\ 0 & 0 & e^{-t} \end{bmatrix} \quad ; \quad \begin{bmatrix} 1 & a & b \\ 0 & 2 + x^2 & c \\ 0 & 0 & -2 - y^2 \end{bmatrix} \quad ; \quad \begin{bmatrix} 0 & 1 & e^{\boldsymbol{i}\theta} \\ 1 & 0 & \boldsymbol{i} \\ e^{-\boldsymbol{i}\theta} & -\boldsymbol{i} & 0 \end{bmatrix}.$$

@ **6.** Soit $M = (1/2)\begin{bmatrix} 5 & 1 & -2 \\ 1 & 5 & -2 \\ 0 & 0 & 4 \end{bmatrix}$.

a) Quelles sont les valeurs propres des matrices M et $M_t = M - tI_3$, où $t \in \mathbb{R}$? Pour quelles valeurs de t les matrices $(M_t)^n$ tendent-elles vers 0 quand n tend vers l'infini ?

b) Posons $A = M_{5/2}$, $B = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ et $U = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Calculer la limite des itérés du vecteur U par la transformation $X \mapsto AX + B$.

@ **7. Recherche d'un régime stationnaire.** On se place dans des conditions expérimentales où des particules peuvent occuper cinq états $e_1, \ldots, e_5$. Les états e_4 et e_5 sont stables. Toute particule occupant l'état e_1 ou e_3 bascule dans l'un des états e_2, e_4 ou e_5 de manière équiprobable ; de même pour une particule dans l'état e_2 vers les états e_1, e_3 ou e_4.

a) Dessiner un graphe où les sommets symbolisent les états et les arcs les changements d'état. Quelle est la matrice M de transition ?

b) Montrer que $M^2 = \begin{bmatrix} \boxed{U} & 0 \\ \boxed{P} & \boxed{I_2} \end{bmatrix}$, où U est une matrice carrée de taille 3 et P une matrice à deux lignes et trois colonnes. Montrer par récurrence que pour tout entier $n \geqslant 1$, on a $M^{2n} = \begin{bmatrix} \boxed{U_n} & 0 \\ \boxed{P_n} & \boxed{I_2} \end{bmatrix}$, où $U_n = \begin{bmatrix} u_n & 0 & u_n \\ 0 & 2u_n & 0 \\ u_n & 0 & u_n \end{bmatrix}$, $P_n = \begin{bmatrix} a_n & b_n & a_n \\ c_n & d_n & c_n \end{bmatrix}$ et $u_{n+1} = \frac{2}{9}u_n$, $a_{n+1} = \frac{8}{9}u_n + a_n$, $b_{n+1} = \frac{10}{9}u_n + b_n$, $c_{n+1} = \frac{2}{3}u_n + c_n$, $d_{n+1} = \frac{4}{9}u_n + d_n$.

c) Calculer u_n et $s_n = u_1 + u_2 + \cdots + u_{n-1}$. En déduire l'expression des coefficients a_n, b_n, c_n et d_n en fonction de n.

d) Montrer que les matrices M^{2n} ont pour limite la matrice

$$M_\infty = \begin{bmatrix} \boxed{\quad 0 \quad} & \boxed{0} \\ \begin{matrix} 4/7 & 5/7 & 4/7 \\ 3/7 & 2/7 & 3/7 \end{matrix} & \boxed{I_2} \end{bmatrix}.$$

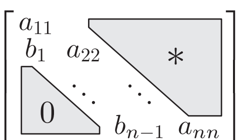
Observer que les vecteurs-colonne de M_∞ sont des vecteurs propres de M pour la valeur propre 1. En déduire que l'on a $\lim_{n\mapsto+\infty}(M^{2n+1}) = MM_\infty = M_\infty$ et donc $M_\infty = \lim_{n\mapsto+\infty}(M^n)$.

e) On part d'une situation initiale où toutes les particules occupent de manière équiprobables les états e_1, e_2, e_3. Montrer qu'après un grand nombre de changements d'état, les particules occupent presque toutes les états e_4 et e_5, dont 62% dans l'état e_4 et 38% dans l'état e_5.

8. Quelques règles sur les valeurs propres. Soit M une matrice carrée de taille n.

a) Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres de M. Montrer que si a est un scalaire, les valeurs propres de aM sont les nombres $a\lambda_i$ et que les valeurs propres de $M + aI_n$ sont les $\lambda_i + a$.

b) Montrer que si λ est valeur propre de M, alors pour tout entier positif q, λ^q est valeur propre de M^q. En déduire qu'une matrice nilpotente n'a que la valeur propre 0.

@ **9. Matrice presque triangulaire.** Soit M une matrice carrée de la forme 

où les coefficients $b_1, \ldots, b_{n-1}$, situés juste sous la diagonale, sont tous non nuls.

a) Montrer que le rang de M est n ou $n-1$ (considérer la sous-matrice obtenue en supprimant la première ligne et la dernière colonne).

b) En déduire que les sous-espaces propres de M sont de dimension 1 (appliquer **(a)** à la matrice $M - \lambda I_n$).

@ **10. Une suite de variables aléatoires binomiales.** Une éprouvette contient trois bactéries, dont une de type A et deux de type B. On les laisse se reproduire en très grand nombre, la proportion de bactéries de chaque type restant inchangée. On prélève alors trois bactéries au hasard que l'on met en culture dans une autre éprouvette. On répète plusieurs fois ces opérations et l'on veut étudier le nombre X_n de bactéries de type A dans l'échantillon qui a été prélevé dans la $(n-1)$-ième culture (la culture initiale porte le numéro 0).

a) Montrer que la probabilité pour qu'on ait prélevé h bactéries de type A dans l'éprouvette initiale est $P(X_1{=}h) = \binom{3}{h}(1/3)^h(2/3)^{3-h}$, où h vaut $0, 1, 2$ ou 3.

b) Soit k l'un des entiers $0, 1, 2$ ou 3. Montrer que la probabilité pour que $X_{n+1} = k$ sachant que $X_n = h$ est $\binom{3}{k}(h/3)^k(1-h/3)^{3-k}$. En déduire l'égalité

$$P(X_{n+1}{=}k) = \sum_{h=0}^{3} \binom{3}{k}(h/3)^k(1-h/3)^{3-k}P(X_n{=}h)\,.$$

c) Notons U_n le vecteur-colonne de $\mathbb{R}^4$ de coordonnées $P(X_n{=}0)$, $P(X_n{=}1)$, $P(X_n{=}2)$, $P(X_n{=}3)$. Montrer que l'on a $U_{n+1} = AU_n$, où $A = \begin{bmatrix} 1 & 8/27 & 1/27 & 0 \\ 0 & 4/9 & 2/9 & 0 \\ 0 & 2/9 & 4/9 & 0 \\ 0 & 1/27 & 8/27 & 1 \end{bmatrix}$.

d) Soit $V = [\,0\ \ 1\ \ 2\ \ 3\,]$. Calculer VA. Montrer que l'espérance de X_n est VU_n. Montrer que cette espérance ne dépend pas de n et qu'en moyenne, on prélève toujours une seule bactérie de type A.

e) Montrer que les vecteurs canoniques $\boldsymbol{E_1}$ et $\boldsymbol{E_4}$ sont vecteurs propres de A. Calculer toutes les valeurs propres de A et les vecteurs propres associés. En déduire que A est diagonalisable.

f) Montrer que l'on a $U_n = A^n U_0$. Quel est le vecteur U_0 ? Calculer A^n. En déduire que la loi de X_n est donnée par

$$P(X_n{=}0) = \frac{2}{3} - \frac{1}{2}\left(\frac{2}{3}\right)^n - \frac{1}{6}\left(\frac{2}{9}\right)^n \ ,\ P(X_n{=}1) = \frac{1}{2}\left[\left(\frac{2}{3}\right)^n + \left(\frac{2}{9}\right)^n\right]$$

$$P(X_n{=}2) = \frac{1}{2}\left[\left(\frac{2}{3}\right)^n - \left(\frac{2}{9}\right)^n\right] \ ,\ P(X_n{=}3) = \frac{1}{3} - \frac{1}{2}\left(\frac{2}{3}\right)^n + \frac{1}{6}\left(\frac{2}{9}\right)^n$$

g) On note F_n la proportion de bactéries de type A dans l'éprouvette numéro n. Montrer que $F_n = X_n/3$. En déduire les probabilités pour que F_n soit égal à 0, à 1/3, à 2/3, à 1. Montrer qu'il y a une probabilité 1/3 pour qu'après un grand nombre de prélèvements, il n'y ait que des bactéries de type A dans l'éprouvette.

@11. Un exemple de système contrôlé. En Biologie, de nombreux transferts de substances sont régis par la loi d'action de masse : par unité de temps, la quantité de substance passant d'un compartiment tissulaire c vers un autre est proportionnelle à la quantité présente dans c.
Considérons une substance qui diffuse dans trois compartiments tissulaires. En notant x, y, z les quantités de substance dans chaque compartiment, on suppose que l'évolution journalière du vecteur $X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ est donnée par $X' - X = MX$, où $M = \begin{bmatrix} -0{,}8 & 0{,}4 & 0{,}4 \\ 0{,}3 & -1 & 0{,}7 \\ 0{,}5 & 0{,}4 & -0{,}9 \end{bmatrix}$ (les coefficients de M déterminent les diffusions entre les trois compartiments). On mesure les quantités initiales x_0, y_0, z_0 et l'on note $X_n = \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix}$ le vecteur-quantité au bout de n jours.

a) Calculer la matrice A telle que $X_{n+1} = AX_n$. Quels sont ses valeurs propres et ses vecteurs propres ?

b) Quelle sont les limites des quantités x_n, y_n, z_n quand n devient grand ?

c) On peut intervenir chaque jour sur les quantités x, y, z en effectuant des injections médicamenteuses : chaque injection a pour effet d'augmenter x d'une certaine

quantité u qu'on peut choisir, de diminuer y d'autant et d'augmenter z de ku, où k est un coefficient constant.

(i) Le paramètre u sert de variable de contrôle. Montrer que ce système linéaire contrôlé s'écrit $X \mapsto AX + Bu$, où $B = \begin{bmatrix} 1 \\ -1 \\ k \end{bmatrix}$.

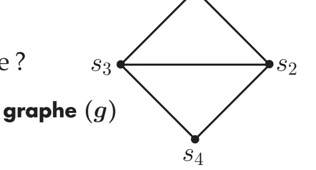
(ii) Calculer la matrice de commandabilité et montrer que si $k = 1$, le système n'est pas commandable.

(iii) Calculer les deux autres valeurs de k pour lesquelles le système n'est pas commandable.

(iv) Le coefficient k vaut $0{,}1$ et l'on a $x_0 = 0{,}17$, $y_0 = 1{,}42$, $z_0 = 0{,}88$. Quelles quantités u_0, u_1, u_2 faut-il utiliser pour obtenir $x_3 = 1{,}2$, $y_3 = 0{,}6$, $z_3 = 0{,}9$?

12. Si G est un graphe de sommets $\{s_1, s_2, \ldots, s_n\}$, sa *matrice d'incidence* est la matrice $M = [m_{ij}]$ carrée de taille n définie par : $m_{ij} = 1$ si les sommets s_i et s_j sont adjacents, $m_{ij} = 0$ sinon.
Si k est un entier positif, un parcours de longueur k dans G est une suite de k sommets adjacents ; notons $m_{ij}(k)$ le coefficient en position i-ème ligne, j-ème colonne dans la matrice M^k.

a) Quelle est la matrice d'incidence du graphe (g) ci-contre ?



graphe (g)

b) Pour un graphe général, montrer que $m_{ij}(2) = m_{i1}m_{1j} + m_{i2}m_{2j} + \cdots + m_{in}m_{nj}$. En déduire que $m_{ij}(2)$ est le nombre de parcours de longueur 2 du sommet s_i au sommet s_j.

c) Montrer que $m_{ij}(k)$ est le nombre de parcours de longueur k du sommet s_i au sommet s_j (raisonner par récurrence).

d) Notons A la matrice d'incidence du graphe (g).

(i) Montrer que A^n est de la forme $\begin{bmatrix} a_n & b_n & b_n & a_n \\ b_n & x_n & y_n & b_n \\ b_n & y_n & x_n & b_n \\ a_n & b_n & b_n & a_n \end{bmatrix}$, où $a_{n+1}=2b_n$ et $b_{n+1}=b_n+4b_{n-1}$.

(ii) De combien de façons peut-on aller de s_1 à s_2 en six pas ? De s_4 à s_4 en sept pas ? De s_2 à s_4 en sept pas ?

Chapitre 7

Espace hermitien, espace euclidien

Nous allons définir dans un cadre assez général les notions de vecteurs orthogonaux et de distance. Rappelons que dans l'espace vectoriel $\mathbb{R}^3$, le produit scalaire des vecteurs $u=(x,y,z)$ et $u'=(x',y',z')$ est défini par $u{\cdot}u'=xx'+yy'+zz'$ et que la norme euclidienne de u est le nombre $\sqrt{u \cdot u} = \sqrt{x^2 + y^2 + z^2}$. Les propriétés du produit scalaire sont

- la symétrie : $u \cdot u' = u' \cdot u$;
- la linéarité par rapport à chaque vecteur : on a en effet

  $(u_1 + u_2) \cdot u' = u_1 \cdot u' + u_2 \cdot u'$

  $(\lambda u) \cdot u' = \lambda(u \cdot u')$ pour tout $\lambda \in \mathbb{R}$

  ce qui assure, par symétrie, la linéarité par rapport au deuxième vecteur ;
- la positivité : $u \cdot u > 0$ si u n'est pas le vecteur nul.

Certaines applications se traitent de manière naturelle dans un $\mathbb{C}$-espace vectoriel : c'est pourquoi nous formulerons les définitions dans le cadre général d'un $\mathbb{K}$-espace vectoriel, où $\mathbb{K}$ désigne $\mathbb{R}$ ou $\mathbb{C}$.

1. Produit hermitien et produit scalaire

Dans ce paragraphe, V est un $\mathbb{K}$-espace vectoriel. Si z est un élément de $\mathbb{K}$, $\bar{z}$ désigne le conjugué de z : on a donc $z=\bar{z} \Longleftrightarrow z \in \mathbb{R}$; rappelons que le module de z, c'est-à-dire la valeur absolue si z est réel, est le nombre réel positif ou nul $|z|$ tel que $|z|^2 = z\bar{z}$.

Définitions

Un *produit hermitien* sur V est la donnée pour tous vecteurs u et u' appartenant à V, d'un nombre $u \cdot u' \in \mathbb{K}$, avec les propriétés suivantes :

i) « symétrie hermitienne » : $u \cdot u' = \overline{u' \cdot u}$

ii) linéarité par rapport au premier vecteur :

$(u_1 + u_2) \cdot u' = u_1 \cdot u' + u_2 \cdot u'$

$(\lambda u) \cdot u' = \lambda(u \cdot u')$ pour tout $\lambda \in \mathbb{K}$

iii) « semi-linéarité » par rapport au second vecteur :

$$u \cdot (u'_1 + u'_2) = u \cdot u'_1 + u \cdot u'_2 \quad \text{et} \quad u \cdot (\lambda u') = \bar{\lambda}(u \cdot u') \text{ pour tout } \lambda \in \mathbb{K},$$

formules qui résultent de (i) et (ii)

iv) positivité : $u \cdot u$ est un nombre réel strictement positif si u n'est pas le vecteur nul.

Le nombre réel $\sqrt{u \cdot u}$ s'appelle la *norme hermitienne* de u et se note $\|u\|$.

Si $\mathbb{K} = \mathbb{R}$, alors $u \cdot u'$ est un nombre réel ; on dit que $u \cdot u'$ est un *produit scalaire* et que $\|u\|$ est la *norme euclidienne*.

En faisant $\lambda = 0$ dans (ii) et (iii), on obtient $u \cdot \mathbf{0} = \mathbf{0} \cdot u = 0$ pour tout vecteur u.

Définitions

Un $\mathbb{C}$-espace vectoriel muni d'un produit hermitien s'appelle un *espace hermitien*.
Un $\mathbb{R}$-espace vectoriel muni d'un produit scalaire s'appelle un *espace euclidien*.

L'espace hermitien $\mathbb{C}^n$. On définit un produit hermitien sur $\mathbb{C}^n$ en posant, pour tous vecteurs $u = (z_1, z_2, \ldots, z_n)$ et $u' = (z'_1, z'_2, \ldots, z'_n)$ appartenant à $\mathbb{C}^n$,

$$u \cdot u' = z_1 \overline{z'_1} + z_2 \overline{z'_2} + \cdots + z_n \overline{z'_n}$$

Les propriétés de symétrie hermitienne et de linéarité par rapport au premier vecteur sont vérifiées ; de plus, le nombre

$$u \cdot u = z_1 \overline{z_1} + z_2 \overline{z_2} + \cdots + z_n \overline{z_n} = |z_1|^2 + |z_2|^2 + \cdots + |z_n|^2$$

est strictement positif si les z_i ne sont pas tous nuls.

Nous dirons que $\mathbb{C}^n$, muni de ce produit hermitien, est *l'espace hermitien $\mathbb{C}^n$ usuel*.

Pour des vecteurs-colonne U et U' appartenant à $\mathbb{C}^n$, on a l'expression matricielle

$$U \cdot U' = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} \begin{bmatrix} \overline{z'_1} \\ \overline{z'_2} \\ \vdots \\ \overline{z'_n} \end{bmatrix} = ({}^tU)\overline{U'},$$

où $\overline{U'}$ s'obtient en conjuguant les coefficients de U'.

L'espace euclidien $\mathbb{R}^n$. Dans l'espace vectoriel $\mathbb{R}^n$, la même formule

$$u \cdot u' = x_1x'_1 + x_2x'_2 + \cdots + x_nx'_n \quad \text{si } u = (x_1, x_2, \ldots, x_n) \text{ et } u' = (x'_1, x'_2, \ldots, x'_n)$$

définit un produit scalaire : on dit que $\mathbb{R}^n$, muni de ce produit scalaire, est *l'espace euclidien* $\mathbb{R}^n$ *usuel* ; la norme de u est $\|u\| = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$.
Pour des vecteurs-colonne U et U' de $\mathbb{R}^n$, le produit scalaire usuel s'écrit

$$U \cdot U' = ({}^tU)U' = ({}^tU')U\,.$$

- On a $\|u - u'\|^2 = (x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2$, donc $|x_i - x'_i| \leqslant \|u - u'\|$ pour tout i.
- Il s'ensuit que si les vecteurs u et u' sont tels que $\|u - u'\| < \varepsilon$, alors leurs coordonnées diffèrent au plus de ε. La norme de $u - u'$ est donc une mesure de l'écart entre les vecteurs u et u'.

L'espace euclidien $\mathbb{R}^3$ est l'espace de la géométrie euclidienne ordinaire.

Exemples

- Dans l'espace euclidien $\mathbb{R}^2$, la norme d'un vecteur (a,b) est $\sqrt{a^2 + b^2}$: les vecteurs de norme 1 sont donc de la forme $(\cos\theta, \sin\theta)$ (figure 1).
- Dans l'espace euclidien $\mathbb{R}^3$, un vecteur (a,b,c) est de norme 1 si l'on a $a^2+b^2+c^2=1$, donc $a^2 + b^2 = \cos^2\varphi$ et $c = \sin\varphi$: les vecteurs de norme 1 sont de la forme $(\cos\theta\cos\varphi, \sin\theta\cos\varphi, \sin\varphi)$, où φ est entre $-\pi/2$ et $\pi/2$.

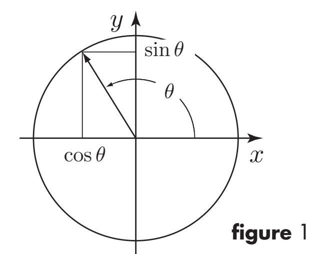


figure 1

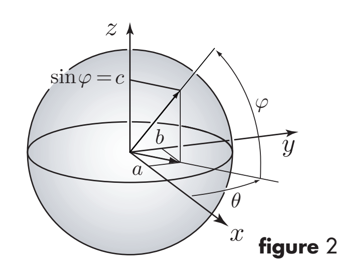


figure 2

Définissons maintenant la notion de vecteurs orthogonaux.

> **Définition**
> Soient u et v des vecteurs d'un espace hermitien ou euclidien. On dit que u et v sont *orthogonaux* si $u \cdot v = 0$.

- Le vecteur nul est orthogonal à tous les vecteurs de V.
- Il n'y a que le vecteur nul qui est orthogonal à lui-même : en effet, d'après la positivité du produit hermitien, l'égalité $u \cdot u = 0$ implique $u = \mathbf{0}$.
- Si un vecteur u est orthogonal à des vecteurs $v_1, v_2, \ldots, v_k$, il est orthogonal à toute combinaison linéaire de $v_1, v_2, \ldots, v_k$, d'après la linéarité du produit hermitien.

Produits scalaires dans $\mathbb{R}^2$

Pour tous vecteurs $X = \begin{bmatrix} x \\ y \end{bmatrix}$ et $X' = \begin{bmatrix} x' \\ y' \end{bmatrix}$ de $\mathbb{R}^2$, posons

$$X \cdot X' = pxx' + q(xy' + x'y) + ryy', \text{ où } p, q, r \text{ sont des nombres réels donnés.}$$

Cette expression est symétrique en X, X', autrement dit $X \cdot X' = X' \cdot X$; de plus, les valeurs dépendent linéairement du couple (x,y), donc pour X' fixé, l'application $X \mapsto X \cdot X'$ est linéaire. Pour que $X \cdot X'$ définisse un produit scalaire, il faut encore que l'expression $X \cdot X = px^2 + 2qxy + ry^2$ ne prenne que des valeurs strictement positives, sauf si $x = y = 0$. Cela exige que ni x, ni y ne soit en facteur, donc p et r doivent être différents de 0. Puisqu'on a alors

$$px^2 + 2qxy + ry^2 = p\left[\left(x + \frac{q}{p}y\right)^2 + \frac{rp - q^2}{p^2}y^2\right]$$

$X \cdot X$ sera strictement positif pour tout $X \neq 0$ si et seulement si $p > 0$ et $rp - q^2 > 0$, c'est-à-dire si $p > 0$ et si le discriminant $q^2 - rp$ de $px^2 + 2qxy + ry^2$ est négatif.

Supposons $p > 0$ et $q^2 - rp < 0$.

L'expression $X \cdot X'$ est alors un produit scalaire sur $\mathbb{R}^2$. On a

$$pxx' + q(xy' + x'y) + ryy' = x(px' + qy') + y(qx' + ry') = [\,x \;\; y\,] \begin{bmatrix} px' + qy' \\ qx' + ry' \end{bmatrix},$$

donc

$$X \cdot X' = [\,x \;\; y\,] \begin{bmatrix} p & q \\ q & r \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}.$$

La matrice $A = \begin{bmatrix} p & q \\ q & r \end{bmatrix}$ est symétrique ; son déterminant $pr - q^2$ est par hypothèse non nul, donc A est inversible.

Norme d'un vecteur. Par définition, on a $\|X\| = \sqrt{X \cdot X} = \sqrt{px^2 + 2qxy + ry^2}$.

Soit M le point de coordonnées (x,y). Pour que le vecteur $\overrightarrow{OM}$ soit de norme 1, il faut et il suffit que M soit sur la courbe d'équation $px^2 + 2qxy + ry^2 = 1$: cette courbe est une ellipse centrée à l'origine (page 219).

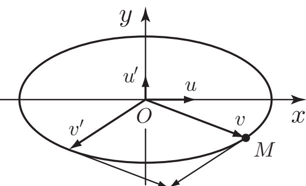
Vecteurs orthogonaux à un vecteur non nul (a,b) donné. L'orthogonalité des vecteurs (x,y) et (a,b) s'exprime par $0 = [\,x \;\; y\,] \begin{bmatrix} p & q \\ q & r \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = x(pa+qb) + y(qa+rb)$.

La matrice A est inversible et le vecteur (a,b) est supposé non nul, donc le vecteur $A\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} pa + qb \\ qa + rb \end{bmatrix}$ n'est pas nul non plus. L'équation $x(pa+qb) + y(qa+rb) = 0$ représente donc une droite vectorielle du plan : c'est la droite vectorielle orthogonale au vecteur (a,b) pour notre produit scalaire.

Exemple. Prenons $A = \begin{bmatrix} p & 0 \\ 0 & r \end{bmatrix}$, où p et r sont des nombres strictement positifs. La matrice A définit le produit scalaire $X \cdot X' = pxx' + ryy'$.

► Les points $M = (x, y)$ tels que $\| \overrightarrow{OM} \| = 1$ sont ceux de l'ellipse d'équation $px^2 + ry^2 = 1$, qu'on peut paramétrer en posant $x = \dfrac{\cos\theta}{\sqrt{p}}$ et $y = \dfrac{\sin\theta}{\sqrt{r}}$, pour $0 \leqslant \theta < 2\pi$. Les axes de l'ellipse sont les axes de coordonnées.

► Soient (a, b) et (a', b') des vecteurs non nuls. Si a et a' sont différents de 0, ces vecteurs ont des pentes $t = b/a$ et $t' = b'/a'$. La condition pour que (a,b) et (a',b') soient orthogonaux est $paa' + rbb' = 0$, ou encore $tt' = -p/r$. Si $a = 0$, la condition d'orthogonalité est $b' = 0$: tout vecteur de l'axe des abscisses est donc orthogonal à tout vecteur de l'axe des ordonnées (les axes de coordonnées sont orthogonaux à la fois pour le produit scalaire usuel et pour le produit scalaire $pxx' + qyy'$).
La figure montre l'ellipse d'équation $x^2 + 4y^2 = k^2$, où $k \neq 0$. Les vecteurs v et v' sont orthogonaux et de norme k pour le produit scalaire $X \cdot X' = xx' + 4yy'$. Si M est un point de l'ellipse, la direction orthogonale à $\overrightarrow{OM}$ est celle de la tangente en M à l'ellipse.



1.1 Calculs dans un espace hermitien

Proposition. *Soient u et v des vecteurs d'un espace hermitien V.*

i) $\|u + v\|^2 = \|u\|^2 + 2\,\mathcal{R}e(u \cdot v) + \|v\|^2$.

ii) $\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2)$ *(formule de la médiane).*

iii) *Si u et v sont orthogonaux, alors $\|u + v\|^2 = \|u\|^2 + \|v\|^2$ (théorème de Pythagore).*

iv) *Pour tout scalaire λ, on a $\|\lambda u\| = |\lambda|\|u\|$.*

Démonstration. Pour montrer la première égalité, développons $\|u+v\|^2 = (u+v)\cdot(u+v)$ en utilisant la définition du produit hermitien. Il vient

$$\begin{aligned}(u+v)\cdot(u+v) = u\cdot(u+v) + v\cdot(u+v) &= u\cdot u + u\cdot v + v\cdot u + v\cdot v \quad \text{par linéarité}\\ &= \|u\|^2 + u\cdot v + \overline{u\cdot v} + \|v\|^2 \quad \text{car } v\cdot u = \overline{u\cdot v}\\ &= \|u\|^2 + 2\,\mathcal{R}e(u\cdot v) + \|v\|^2\,.\end{aligned}$$

La seconde formule s'en déduit, car $u\cdot(-v) = -(u\cdot v)$. Si u et v sont orthogonaux, alors $u\cdot v = 0$ et la troisième formule n'est autre que (i). Enfin, on a $(\lambda u)\cdot(\lambda u) = \lambda\bar{\lambda}(u\cdot u)$, c'est-à-dire $\|\lambda u\|^2 = |\lambda|^2\|u\|^2$, d'où (iv) puisque norme et module sont des réels positifs ou nuls. ■

Remarque

Dans un espace euclidien, l'égalité (i) s'écrit simplement $\|u+v\|^2 = \|u\|^2 + 2u\cdot v + \|v\|^2$, car le produit scalaire est un nombre réel.

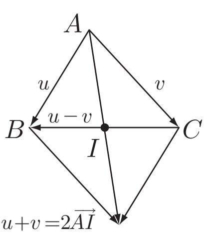
Interprétation géométrique

► L'égalité (ii) s'appelle la formule de la médiane car elle possède une interprétation simple en géométrie :

si A, B, C sont des points de l'espace euclidien usuel et si I est le milieu de B, C, alors

$$AB^2 + AC^2 = 2AI^2 + (1/2)BC^2 \,.$$

En posant $u = \overrightarrow{AB}$ et $v = \overrightarrow{AC}$, on a en effet $u + v = 2\overrightarrow{AI}$, $u - v = \overrightarrow{CB}$, $AB = \|u\|$, $AC = \|v\|$ et d'après (ii), il vient $(2AI)^2 + CB^2 = 2(AB^2 + AC^2)$.



► La formule (iii) généralise le théorème de Pythagore : en effet, si le triangle ABC est rectangle en A, les vecteurs $\overrightarrow{BA}$ et $\overrightarrow{AC}$ sont orthogonaux et puisque $\overrightarrow{BA} + \overrightarrow{AC} = \overrightarrow{BC}$, l'égalité (iii) s'écrit $BA^2 + AC^2 = BC^2$.

Inégalité de Cauchy-Schwarz. *Pour tous vecteurs u,v de V, on a $|u\cdot v| \leqslant \|u\|\,\|v\|$. Si u et v sont non nuls et si on a l'égalité, alors u et v sont colinéaires.*

Démonstration. Si $u \cdot v = 0$, l'inégalité est vraie. Supposons $u \cdot v \neq 0$. Soit z le nombre complexe tel que $z(u\cdot v) = |u \cdot v|$; ce nombre z est donc de module 1 et vérifie $(zu)\cdot v = |u\cdot v|$. Pour tout nombre réel t, on a

$$\begin{aligned}\|zu + tv\|^2 &= \|zu\|^2 + 2\,\mathcal{R}e(zu \cdot tv) + \|tv\|^2\\ &= \|u\|^2 + 2t\,\mathcal{R}e(zu \cdot v) + t^2\|v\|^2 \quad \text{car } t \text{ est réel et } \|zu\| = |z|\|u\| = \|u\|.\\ &= \|u\|^2 + 2t|u \cdot v| + t^2\|v\|^2 \quad \text{par définition de } z.\end{aligned}$$

La fonction $t \mapsto t^2\|v\|^2 + 2t|u \cdot v| + \|u\|^2$ ne prend que des valeurs positives ou nulles, donc le discriminant réduit $|u \cdot v|^2 - \|u\|^2\,\|v\|^2$ est négatif ou nul, d'où l'inégalité $|u \cdot v| \leqslant \|u\|\,\|v\|$. Considérons le vecteur $w = \|v\|^2u - (u \cdot v)v$. On a

$$\begin{aligned}\|w\|^2 &= \|v\|^4\|\,\|u\|^2 - 2\,\mathcal{R}e\big(\|v\|^2u \cdot (u \cdot v)v\big) + |u \cdot v|^2\|v\|^2\\ &= \|v\|^4\,\|u\|^2 - 2\,\mathcal{R}e\big(\|v\|^2(\overline{u \cdot v})(u \cdot v)\big) + |u \cdot v|^2\|v\|^2\\ &= \|v\|^4\,\|u\|^2 - 2\|v\|^2|u \cdot v|^2 + |u \cdot v|^2\|v\|^2 = \|v\|^2\big(\|u\|^2\|v\|^2 - |u \cdot v|^2\big)\,.\end{aligned}$$

Supposons qu'on a l'égalité $|u\cdot v| = \|u\|\,\|v\|$. Alors il vient $\|w\|^2 = 0$, donc $w = 0$ et les vecteurs u et v sont colinéaires. ■

Propriétés de la norme. *Pour tous vecteurs u et v appartenant à V et pour tout scalaire $\lambda \in \mathbb{K}$, on a*

i) $\|u\| \geqslant 0$ *et l'équivalence* $(\|u\| = 0 \Longleftrightarrow u = 0)$
ii) $\|\lambda u\| = |\lambda|\|u\|$
iii) $\|u + v\| \leqslant \|u\| + \|v\|$ *(inégalité triangulaire).*

Démonstration. Nous avons déjà remarqué les propriétés (i) et (ii). Montrons l'inégalité (iii). On a $\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2\,\mathcal{R}e(u\cdot v)$. La partie réelle d'un nombre complexe est inférieure ou égale au module, donc il vient

$$\|u + v\|^2 \leqslant \|u\|^2 + \|v\|^2 + 2|u \cdot v| \leqslant \|u\|^2 + \|v\|^2 + 2\|u\|\,\|v\|$$

d'après l'inégalité de Cauchy-Schwarz. On a ainsi $\|u+v\|^2 \leqslant (\|u\| + \|v\|)^2$, d'où le résultat. ■

La norme d'un vecteur u mesure l'écart entre u et le vecteur nul ; la norme du vecteur $u - v$ est une distance entre u et v.

1.2 Base orthonormée

Dans un espace hermitien, les combinaisons linéaires de vecteurs deux à deux orthogonaux conduisent à des calculs plus simples.

Proposition.

- *Si $u_1, u_2, \ldots, u_n$ sont des vecteurs de V deux à deux orthogonaux, alors $\|u_1+u_2+\cdots+u_n\|^2 = \|u_1\|^2 + \|u_2\|^2 + \cdots + \|u_n\|^2$ (formule de Pythagore).*
- *Des vecteurs non nuls et deux à deux orthogonaux sont indépendants.*

Démonstration. La première propriété se démontre par récurrence sur le nombre de vecteurs, en utilisant le théorème de Pythagore (page 203). Supposons que les vecteurs $u_1, \ldots, u_n$ sont deux à deux orthogonaux et tous non nuls, et soient x_i des scalaires. Pour tout $p = 1, 2, \cdots, n$, on a par linéarité

$$(x_1u_1 + x_2u_2 + \cdots + x_nu_n) \cdot u_p = x_1(u_1 \cdot u_p) + x_2(u_2 \cdot u_p) + \cdots + x_n(u_n \cdot u_p) = x_p(u_p \cdot u_p)$$

car $u_i \cdot u_p = 0$ si $i \neq p$. Si $x_1u_1 + x_2u_2 + \cdots + x_nu_n = 0$, alors $x_p(u_p \cdot u_p) = 0$. On en déduit $x_p = 0$, car, u_p n'étant pas le vecteur nul, on a $u_p \cdot u_p \neq 0$. ■

Supposons que V est de dimension finie.

Définition

Une base $e_1, e_2, \ldots, e_n$ de V est dite *orthonormée* si les vecteurs e_i sont deux à deux orthogonaux et de norme 1.

D'après la proposition, pour que des vecteurs $e_1, e_2, \ldots, e_n$ forment une base orthonormée d'un espace hermitien V de dimension n, il faut et il suffit qu'ils appartiennent à V, qu'ils soient deux à deux orthogonaux et tous de norme 1.

Exemples

- La base canonique de $\mathbb{R}^n$ est une base orthonormée pour le produit scalaire usuel.
- Les vecteurs $\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$ forment une base orthonormée de l'espace euclidien $\mathbb{R}^2$.
- Les vecteurs $e_1 = \begin{bmatrix} \cos\theta\cos\varphi \\ \sin\theta\cos\varphi \\ \sin\varphi \end{bmatrix}$, $e_2 = \begin{bmatrix} \cos\theta\sin\varphi \\ \sin\theta\sin\varphi \\ -\cos\varphi \end{bmatrix}$, $e_3 = \begin{bmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{bmatrix}$ forment une base orthonormée de $\mathbb{R}^3$.

Remarque

Si e est un vecteur non nul de V, le vecteur $e' = (1/\|e\|)e$ est de norme 1, car d'après les propriétés de la norme, on a $\|e'\| = (1/\|e\|)\|e\| = 1$. Si des vecteurs sont non nuls et deux à deux orthogonaux, alors en les divisant par leur norme, on obtient des vecteurs deux à deux orthogonaux et de norme 1.

Proposition. *Si $e_1, e_2, \ldots, e_n$ est une base orthonormée de V, alors pour tout vecteur $v \in V$, on a $v = (v \cdot e_1)e_1 + (v \cdot e_2)e_2 + \cdots + (v \cdot e_n)e_n$.*

Démonstration. Soient $x_1, x_2, \ldots, x_n$ les coordonnées de v dans la base $e_1, e_2, \ldots, e_n$, de sorte que $v = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$. Pour tout $p = 1, 2, \ldots, n$, on a

$$v \cdot e_p = (x_1 e_1) \cdot e_p + (x_2 e_2) \cdot e_p + \cdots + (x_n e_n) \cdot e_p = x_1(e_1 \cdot e_p) + x_2(e_2 \cdot e_p) + \cdots + x_n(e_n \cdot e_p)$$

Puisque $e_i \cdot e_p = 0$ pour $i \neq p$, il vient $v \cdot e_p = x_p(e_p \cdot e_p) = x_p$, car $e_p \cdot e_p = 1$. ■

Les coordonnées d'un vecteur dans une base orthonormée s'obtiennent donc simplement en calculant les produits hermitiens avec les vecteurs de la base. Il est également aisé de calculer le produit hermitien de deux vecteurs dont on connaît les coordonnées dans une base orthonormée.

Proposition. *Si des vecteurs u et u' ont pour coordonnées $(x_1, x_2, \ldots, x_n)$ et $(x'_1, x'_2, \ldots, x'_n)$ dans une base orthonormée de V, alors*

$$u \cdot u' = x_1 \overline{x'_1} + x_2 \overline{x'_2} + \cdots + x_n \overline{x'_n} \qquad \text{et} \qquad \|u\|^2 = |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2.$$

Démonstration. Notons $e_1, e_2, \ldots e_n$ la base orthonormée considérée. On a $e_i \cdot u' = \overline{u' \cdot e_i} = \overline{x'_i}$, d'après la proposition précédente. Puisque $u = x_1 e_1 + \cdots + x_n e_n$, on en déduit par linéarité la formule pour le produit hermitien $u \cdot u'$. ■

Construction de bases orthonormées

Voici un moyen d'obtenir une base orthonormée à partir d'une base quelconque de l'espace vectoriel V. Cette méthode s'appelle *l'algorithme de Gram-Schmidt*.

Supposons par exemple V de dimension 3 et soit v_1, v_2, v_3 une base de V.

Étape 1. On pose $e_1 = \dfrac{1}{\|v_1\|} v_1$, ce qui est possible car le vecteur v_1 n'est pas nul. Le vecteur e_1 est de norme 1.

Étape 2. Dans le plan engendré par v_1, v_2, on considère un vecteur $v = \lambda e_1 + v_2$ et l'on cherche le scalaire λ pour que v soit orthogonal à e_1. La condition s'écrit

$$\begin{aligned} 0 = v \cdot e_1 &= (\lambda e_1 + v_2) \cdot e_1 \\ &= \lambda e_1 \cdot e_1 + v_2 \cdot e_1 = \lambda + v_2 \cdot e_1 \text{ car } e_1 \cdot e_1 = 1 \end{aligned}$$

Il vient $\lambda = -v_2 \cdot e_1$ d'où $v = -(v_2 \cdot e_1)e_1 + v_2$.
Puisque v_1 et v_2 sont indépendants, e_1 et v_2 le sont aussi, donc v n'est pas nul. On pose alors $e_2 = \dfrac{1}{\|v\|} v$: ainsi e_2 est de norme 1 et orthogonal à e_1.

Étape 3. On cherche ensuite un vecteur $w = \lambda_1 e_1 + \lambda_2 e_2 + v_3$ orthogonal à e_1 et à e_2 :

$$0 = w \cdot e_1 = \lambda_1 e_1 \cdot e_1 + \lambda_2 e_2 \cdot e_1 + v_3 \cdot e_1 = \lambda_1 + v_3 \cdot e_1 \quad \text{car } e_2 \cdot e_1 = 0 \text{ et } e_1 \cdot e_1 = 1$$

$$0 = w \cdot e_2 = \lambda_1 e_1 \cdot e_2 + \lambda_2 e_2 \cdot e_2 + v_3 \cdot e_2 = \lambda_2 + v_3 \cdot e_2 \quad \text{car } e_1 \cdot e_2 = 0 \text{ et } e_2 \cdot e_2 = 1$$

Le vecteur $w = -(v_3 \cdot e_1)e_1 - (v_3 \cdot e_2)e_2 + v_3$, est donc orthogonal à e_1 et à e_2. Puisque v_3 n'est pas combinaison linéaire de v_1 et v_2, il n'est pas non plus combinaison linéaire de e_1 et e_2, donc w n'est pas nul. On pose $e_3 = \dfrac{1}{\|w\|} w$: le vecteur e_3 est de norme 1 et orthogonal à e_1 et e_2.

Les vecteurs e_1, e_2, e_3 forment ainsi une base orthonormée de V.

Cet algorithme montre que : *tout espace hermitien de dimension finie possède des bases orthonormées*.

1.3 Un exemple d'espace hermitien

Notons V l'ensemble des fonctions continues sur $[0,T]$ et à valeurs dans $\mathbb{C}$. La somme de deux fonctions continues est continue, ainsi que le produit d'une fonction continue par un nombre, donc V est un $\mathbb{C}$-espace vectoriel.
Définissons le produit hermitien des fonctions en posant

$$f \cdot g = \frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt \ , \quad \text{pour toutes fonctions } f \text{ et } g \text{ appartenant à } V.$$

Vérifions que les propriétés du produit hermitien sont bien satisfaites.
Rappelons que pour intégrer une fonction à valeurs complexes, on intègre sa partie réelle et sa partie imaginaire. Il s'ensuit que l'on a $\int_0^T \overline{f(t)}\,dt = \overline{\int_0^T f(t)\,dt}$.

- $f\cdot g=\frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt=\frac{1}{T}\int_0^T \overline{g(t)\overline{f(t)}}\,dt=\frac{1}{T}\overline{\int_0^T g(t)\overline{f(t)}\,dt}=\overline{g\cdot f}$, d'où la symétrie hermitienne.
- D'après la linéarité de l'intégrale, on a $(\lambda f)\cdot g=\lambda(f\cdot g)$ et $(f_1+f_2)\cdot g=f_1\cdot g+f_2\cdot g$.
- $f\cdot f=\frac{1}{T}\int_0^T f(t)\overline{f(t)}\,dt=\frac{1}{T}\int_0^T |f(t)|^2\,dt$ est un nombre réel positif ou nul. Comme la fonction $x\mapsto|f(x)|^2$ est continue et positive ou nulle sur $[0,T]$, cette intégrale n'est nulle que si $|f(x)|=0$ pour tout x, c'est-à-dire si f est la fonction nulle (page 288).

La norme d'une fonction $f\in V$ est $\|f\| = \left(\frac{1}{T}\int_0^T |f(t)|^2\,dt\right)^{1/2}$ et l'inégalité de Cauchy-Schwarz s'écrit (en multipliant par T)

$$(*) \qquad \left|\int_0^T f(t)\overline{g(t)}\,dt\right| \leqslant \left(\int_0^T |f(t)|^2\,dt\right)^{1/2}\left(\int_0^T |g(t)|^2\,dt\right)^{1/2}$$

Une famille de vecteurs orthogonaux. Posons $\omega = 2\pi/T$. Pour tout entier $n\in\mathbb{Z}$, définissons la fonction $e_n:\mathbb{R}\to\mathbb{C}$ par la formule

$$e_n(x) = e^{n\boldsymbol{i}\omega x} \quad , \quad \text{pour tout } x\in\mathbb{R}.$$

Ces fonctions e_n sont continues et comme on a $n\boldsymbol{i}\omega(x+T)=n\boldsymbol{i}\omega x+2\pi n\boldsymbol{i}$, il vient $e_n(x+T)=e_n(x)$: les fonctions e_n sont donc périodiques de période T.
Si $n\neq 0$, il s'ensuit $\int_0^T e^{n\boldsymbol{i}\omega t}\,dt = \frac{1}{n\boldsymbol{i}\omega}\left[e^{n\boldsymbol{i}\omega t}\right]_0^T = 0$. Puisque $e_0(x)=1$, il vient

$$\int_0^T e_n(t)\,dt = 0 \text{ si } n\neq 0 \quad \text{et} \quad \frac{1}{T}\int_0^T e_0(t)\,dt = 1$$

Remarquons les formules $\overline{e_p(x)} = e^{-p\boldsymbol{i}\omega x} = e_{-p}(x)$ et $e_n(x)e_p(x) = e^{n\boldsymbol{i}\omega x}e^{p\boldsymbol{i}\omega x} = e_{n+p}(x)$. On en déduit que pour tous entiers relatifs n et p, on a

$$\int_0^T e_n(t)\overline{e_p(t)}\,dt = \int_0^T e_n(t)e_{-p}(t)\,dt = \int_0^T e_{n-p}(t)\,dt$$

donc

$$e_n \cdot e_p = \begin{cases} 0 & \text{si } n \neq p \\ 1 & \text{si } n = p \end{cases}$$

Dans l'espace V, les vecteurs e_n sont deux à deux orthogonaux et de norme 1. En particulier, pour tout entier $k \geqslant 1$, les vecteurs $e_0, e_1, \ldots, e_k$ sont indépendants, donc l'espace vectoriel V n'est pas de dimension finie.

Le produit hermitien qu'on vient de définir joue un rôle important lorsqu'on veut approcher un signal périodique par des fonctions trigonométriques (voir page 549).

1.4 Sous-espace orthogonaux et projections

Définition
Soit W un sous-espace vectoriel d'un espace hermitien V. On dit qu'un vecteur $u \in V$ est *orthogonal à* W s'il est orthogonal à tous les vecteurs de W.

Le vecteur nul est toujours orthogonal à W. Si u et u' sont des vecteurs orthogonaux à W, alors on a $u \cdot w = 0 = u' \cdot w$ pour tout $w \in W$, donc aussi $(\lambda u) \cdot w = 0$ et $(u + u') \cdot w = 0$: les vecteurs λu et $u + u'$ sont donc orthogonaux à W. Ainsi l'ensemble des vecteurs de V qui sont orthogonaux à W est un sous-espace vectoriel de V.

Définition
Soit W un sous-espace vectoriel de V. Le sous-espace vectoriel de V formé des vecteurs orthogonaux à W s'appelle *l'orthogonal de* W et se note $W^\perp$.

Proposition. *Les sous-espaces W et $W^\perp$ n'ont en commun que le vecteur nul.*

En effet, si u est dans $W^\perp$, il est orthogonal à tous les vecteurs de W ; si de plus $u \in W$, alors u est en particulier orthogonal à u, donc u est le vecteur nul.

Attention : lorsque W n'est pas de dimension finie, $W^\perp$ peut être réduit au vecteur nul (exercice 11).

Projection sur un sous-espace de dimension finie

Supposons que W est un sous-espace de dimension finie de V.

Il existe donc une base orthonormée $e_1, e_2, \ldots, e_n$ de W. Soit $u \in V$ et soit $u' = (u \cdot e_1)e_1 + (u \cdot e_2)e_2 + \cdots + (u \cdot e_n)e_n$. Le vecteur u' appartient à W, ses coordonnées dans la base $e_1, e_2, \ldots, e_n$ sont les $u \cdot e_i$, donc $u' \cdot e_i = u \cdot e_i$ d'après une proposition page 205. Il s'ensuit $(u - u') \cdot e_i = 0$ pour tout i, donc $u - u'$ est orthogonal aux vecteurs $e_1, e_2, \ldots, e_n$. Comme tout vecteur de W est combinaison linéaire de $e_1, e_2, \ldots, e_n$, le vecteur $u - u'$ est orthogonal à W, autrement dit $u - u' \in W^\perp$.

Théorème de la projection. *Soit W un sous-espace vectoriel de dimension finie de V.*

- *Pour tout vecteur $u \in V$, il existe un unique vecteur $p_W(u) \in W$ tel que $u - p_W(u) \in W^\perp$. Le vecteur $p_W(u)$ s'appelle le projeté orthogonal de u sur W.*
- *Si $e_1, e_2, \ldots, e_n$ est une base orthonormée de W, alors*

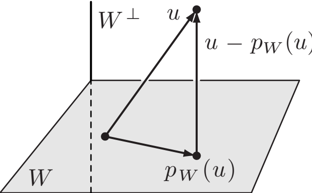
$$p_W(u) = (u \cdot e_1)e_1 + (u \cdot e_2)e_2 + \cdots + (u \cdot e_n)e_n$$

- *L'application $p_W : V \to W$ est linéaire et s'appelle la projection orthogonale sur W.*
- *Si u est un vecteur de V, alors*
  - i) $\|u\|^2 = \|p_W(u)\|^2 + \|u - p_W(u)\|^2$
  - ii) $u \in W \iff p_W(u) = u$ *et* $u \in W^\perp \iff p_W(u) = \mathbf{0}$
  - iii) *pour tout vecteur $w \in W$, on a $\|u - w\| \geqslant \|u - p_W(u)\|$.*

Démonstration. Nous avons montré ci-dessus que si $e_1, e_2, \ldots, e_n$ est une base orthonormée de W, alors en posant

$$p_W(u) = u' = (u \cdot e_1)e_1 + (u \cdot e_2)e_2 + \cdots + (u \cdot e_n)e_n\,,$$

on a $u' \in W$ et $u - u' \in W^\perp$.



Montrons l'unicité du projeté orthogonal. Supposons que u'' est aussi un vecteur de W tel que $u - u'' \in W^\perp$. Le vecteur $(u - u'') - (u - u')$ appartient à $W^\perp$ et $u' - u''$ appartient à W ; puisqu'on a $(u - u'') - (u - u') = u' - u''$, ce vecteur appartient à W et à $W^\perp$, donc il est nul. D'après l'expression ci-dessus de $p_W(u)$ dans une base orthonormée de W, l'application $u \mapsto p_W(u)$ est linéaire.

La propriété (i) est une application du théorème de Pythagore. Montrons (ii) : si $u \in W$, alors par définition du projeté orthogonal, on a $u = p_W(u)$ car $\mathbf{0} = u - u$ est dans $W^\perp$; réciproquement, si $u = p_W(u)$, alors $u \in W$, car $p_W(u) \in W$; si $u \in W^\perp$, alors $u - \mathbf{0} \in W^\perp$, donc $p_W(u) = \mathbf{0}$; réciproquement, si $p_W(u) = \mathbf{0}$, alors $u - p_W(u) = u$ est dans $W^\perp$.

Soit $u \in V$ et $w \in W$. Puisque $u - p_W(u) \in W^\perp$ et $p_W(u) - w \in W$, ces vecteurs sont orthogonaux. D'après le théorème de Pythagore, il vient

$$\|u - w\|^2 = \|\big(u - p_W(u)\big) + \big(p_W(u) - w\big)\|^2 = \|u - p_W(u)\|^2 + \|p_W(u) - w\|^2 \geqslant \|u - p_W(u)\|^2\,.$$

■

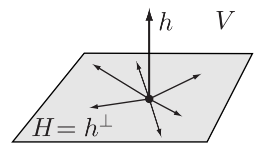
Voici la signification de la propriété (iii) : *si l'on cherche à approcher un vecteur de $u \in V$ par un vecteur de W, la meilleure approximation en norme est le projeté orthogonal $p_W(u)$.*

Cas d'un espace V de dimension finie

Proposition. *Si l'espace hermitien V est de dimension finie, alors pour tout sous-espace vectoriel W de V, on a $\dim W + \dim(W^\perp) = \dim V$ et $(W^\perp)^\perp = W$.*

Démonstration. La projection orthogonale $p_W : V \to W$ est une application linéaire d'image W et de noyau $W^\perp$. D'après la formule de la dimension (page 172), on a donc $\dim W + \dim(W^\perp) = \dim V$. En appliquant cette égalité à $W^\perp$, il vient $\dim\big((W^\perp)^\perp\big) = \dim V - \dim(W^\perp) = \dim W$. Tout vecteur de W est orthogonal à tout vecteur de $W^\perp$, donc on a l'inclusion $W \subset \big(W^\perp\big)^\perp$. D'après la propriété (4) page 165, on en déduit l'égalité $W = \big(W^\perp\big)^\perp$. ■

Hyperplans d'un espace euclidien de dimension finie. Supposons que V est un espace euclidien de dimension n.



- *Si h est un vecteur non nul de V, l'ensemble des vecteurs de V orthogonaux à h est un hyperplan de V, appelé hyperplan orthogonal à h et noté $h^{\perp}$.*
- *Pour tout hyperplan H de V, il existe un vecteur $h \neq 0$ tel que $H = h^{\perp}$.*
- *Soit $\mathscr{B}$ une base orthonormée de V. Un sous-espace H de V est un hyperplan si et seulement s'il a dans $\mathscr{B}$ une équation de la forme $a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$, où les scalaires a_i ne sont pas tous nuls. Dans ce cas, H est l'orthogonal du vecteur de coordonnées $(a_1, a_2, \ldots, a_n)$.*

Démonstration. Si h est un vecteur non nul de V, le sous-espace engendré par h est de dimension 1, donc $h^{\perp}$ est de dimension $n-1$. Réciproquement, si H est un sous-espace de dimension $n-1$, alors $H^{\perp}$ est de dimension 1 : en choisissant un vecteur non nul $h \in H^{\perp}$, on a donc $h^{\perp} = \left(H^{\perp}\right)^{\perp} = H$. Soient $(a_1, a_2, \ldots, a_n)$ les coordonnées de h dans la base $\mathscr{B}$. Un vecteur x de coordonnées $(x_1, x_2, \ldots, x_n)$ est orthogonal à h si et seulement si $a_1x_1 + a_2x_2 + \cdots + a_nx_n = h \cdot x = 0$. ■

Par le théorème de la projection, il est très simple de calculer le projeté d'un vecteur sur un sous-espace W dont on connaît une base orthonormée. On peut cependant avoir besoin d'exprimer le vecteur projeté au moyen d'une base quelconque de W. Avant d'établir une telle formule, démontrons un résultat préliminaire.

Proposition. *Soit A une matrice réelle à n lignes et p colonnes. Si $p \leqslant n$ et si les colonnes de A sont indépendantes, alors la matrice $({}^tA)A$ est carrée de taille p et inversible.*

Démonstration. Supposons que $X \in \mathbb{R}^p$ est un vecteur-colonne tel que $({}^tA\,A)X = 0$. En multipliant à gauche par la matrice-ligne ${}^tX$, il vient $0 = ({}^tX\,{}^tA)(AX) = {}^t(AX)(AX) = \|AX\|^2$, où la norme est la norme euclidienne dans $\mathbb{R}^n$. On a donc $AX = 0$. Puisque les colonnes de A sont indépendantes, on en déduit $X = 0$. Cela montre que l'équation linéaire $({}^tA\,A)X = 0$ a pour seule solution $X = 0$: la matrice carrée ${}^tA\,A$ est donc inversible. ■

Proposition. *Soient $u_1, u_2, \ldots, u_p$ des vecteurs indépendants appartenant à $\mathbb{R}^n$ et soit W le sous-espace vectoriel de $\mathbb{R}^n$ engendré par $u_1, u_2, \ldots, u_p$. Soit A la matrice ayant pour colonnes les coordonnées de $u_1, u_2, \ldots, u_p$. Alors pour tout vecteur-colonne $X \in \mathbb{R}^n$, le projeté orthogonal de X sur W est $p_W(X) = y_1u_1 + \cdots + y_pu_p$, où $\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = ({}^tA\,A)^{-1}({}^tA)X$.*

- Cette formule donne les coordonnées du projeté orthogonal dans la base $u_1, u_2, \ldots, u_p$ de W.

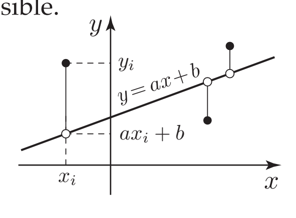
► Si $A_1, A_2, \dots, A_p$ sont les colonnes de A, les coordonnées de $p_W(X)$ dans la base canonique sont $y_1 A_1 + y_2 A_2 + \cdots + y_p A_p = A \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}$.

La projection $p_W : \mathbb{R}^n \to \mathbb{R}^n$ a pour matrice $A({}^tA\,A)^{-1}({}^tA)$ dans la base canonique de $\mathbb{R}^n$.

Démonstration de la proposition. Tout vecteur $X \in \mathbb{R}^n$ se décompose en $X = X' + X''$, où $X' \in W$ et $X'' \in W^\perp$. En notant A_i les colonnes de A, les coefficients de la matrice-colonne $({}^tA)X''$ sont les produits scalaires $({}^tA_i)X'' = u_i \cdot X''$ qui sont nuls puisque X'' est orthogonal à tous les vecteurs de W. Il vient donc $({}^tA\,A)^{-1}({}^tA)X = ({}^tA\,A)^{-1}({}^tA)X' + ({}^tA\,A)^{-1}({}^tA)X'' = ({}^tA\,A)^{-1}({}^tA)X'$. Puisque $X' \in W$, on a $X' = y_1A_1 + y_2A_2 + \cdots + y_pA_p = AY$, où Y est la matrice-colonne des coefficients y_i. Alors $({}^tA\,A)^{-1}({}^tA)X' = ({}^tA\,A)^{-1}({}^tA\,A)Y = I_pY = Y$, d'où le résultat puisque $X' = p_W(X)$ par définition du projeté orthogonal. ■

1.5 Une application : la méthode des moindres carrés

Le problème. Supposons qu'une quantité réelle y dépende d'un paramètre x et qu'on dispose de n mesures $y_1, y_2, \dots, y_n$ correspondant à des valeurs $x_1, x_2, \dots, x_n$ du paramètre ; on suppose $n \geqslant 2$ et que les x_i ne sont pas tous égaux.
Il s'agit de trouver une relation de la forme $y = ax + b$ qui approxime le mieux la relation entre les x_i et les y_i : précisément, on cherche à déterminer a et b pour que la somme des écarts $\left[y_i - (ax_i + b)\right]^2$ soit la plus petite possible.
La droite d'équation $y = ax + b$ sera celle qui s'écarte globalement le moins des points (x_i, y_i) ; on l'appelle la *droite de régression*.



La droite de régression permet de calculer pour tout x une valeur raisonnable de y : on peut ainsi faire des extrapolations.
Pour déterminer la droite de régression, nous utiliserons le théorème de projection dans l'espace euclidien usuel $\mathbb{R}^n$.

Résolution

Notons X le vecteur-colonne de $\mathbb{R}^n$ de coefficients $(x_1, x_2, \dots, x_n)$, Y le vecteur-colonne de coefficients $(y_1, y_2, \dots, y_n)$ et U le vecteur-colonne de $\mathbb{R}^n$ dont toutes les coordonnées sont égales à 1.
Pour tous nombres a et b, les $y_i - ax_i - b$ sont les coordonnées du vecteur $Y - aX - bU$ et la somme des $(y_i - ax_i - b)^2$ est égale à $\|Y - aX - bU\|^2$. Quand a et b parcourent les nombres réels, le vecteur $aX + bU$ parcourt le sous-espace vectoriel $\mathscr{W}$ de $\mathbb{R}^n$ engendré par X et U. Puisque les nombres x_i ne sont pas tous égaux, les vecteurs X et U ne sont pas colinéaires, donc ils forment une base de $\mathscr{W}$.
D'après la propriété (iii) du théorème de projection, la norme $\|Y - aX - bU\|$ sera minimum si $aX + bU$ est la projection orthogonale de Y sur $\mathscr{W}$.

Soit $A = [\, X \;\; U\,]$ la matrice à n lignes dont les deux colonnes sont X et U. D'après la proposition précédente, le projeté orthogonal de Y sur $\mathscr{W}$ est $A({}^tA\,A)^{-1}({}^tA\,Y)$.

► La matrice ${}^tA\,A$ est carrée de taille 2 : on a

$$ {}^tA\,A = \begin{bmatrix} {}^tX \\ {}^tU \end{bmatrix} [\, X \;\; U\,] = \begin{bmatrix} {}^tXX & {}^tXU \\ {}^tUX & {}^tUU \end{bmatrix} = \begin{bmatrix} \|X\|^2 & X \cdot U \\ U \cdot X & \|U\|^2 \end{bmatrix}. $$

Le produit scalaire $X \cdot U$ est la somme $s_x = x_1 + x_2 + \cdots + x_n$ et $\|U\|^2 = n$. En posant $s_{x^2} = x_1^2 + x_2^2 + \cdots + x_n^2 = \|X\|^2$, il vient donc

$$ ({}^tA\,A)^{-1} = \begin{bmatrix} s_{x^2} & s_x \\ s_x & n \end{bmatrix}^{-1} = \frac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} n & -s_x \\ -s_x & s_{x^2} \end{bmatrix} $$

► On a ${}^tA\,Y = \begin{bmatrix} {}^tX \\ {}^tU \end{bmatrix} Y = \begin{bmatrix} {}^tX\,Y \\ {}^tU\,Y \end{bmatrix} = \begin{bmatrix} s_{xy} \\ s_y \end{bmatrix}$, où $s_{xy} = x_1y_1 + x_2y_2 + \cdots + x_ny_n$ et $s_y = y_1 + y_2 + \cdots + y_n$.

Il vient $({}^tA\,A)^{-1}({}^tA\,Y) = \dfrac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} n & -s_x \\ -s_x & s_{x^2} \end{bmatrix} \begin{bmatrix} s_{xy} \\ s_y \end{bmatrix} = \dfrac{1}{ns_{x^2} - s_x^2} \begin{bmatrix} ns_{xy} - s_xs_y \\ s_{x^2}s_y - s_xs_{xy} \end{bmatrix}$.

En posant

$$ a = \frac{ns_{xy} - s_xs_y}{ns_{x^2} - s_x^2} \quad \text{et} \quad b = \frac{s_{x^2}s_y - s_xs_{xy}}{ns_{x^2} - s_x^2}, $$

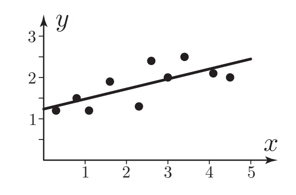
on obtient $p_{\mathscr{W}}(Y) = A \begin{bmatrix} a \\ b \end{bmatrix} = aX + bU$.

La droite de régression a donc pour équation $y = ax + b$.

Par exemple, pour les données du tableau suivant,

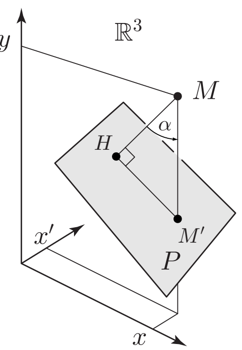
x_i	0,3	0,8	1,1	1,6	2,3	2,6	3,0	3,4	4,1	4,5
y_i	1,2	1,5	1,2	1,9	1,3	2,4	2,0	2,5	2,1	2,0

la droite de régression a pour équation $y = 0{,}24x + 1{,}23$.



Généralisation. La méthode précédente s'applique aussi bien lorsque les données y_i dépendent de plusieurs paramètres $x, x', \ldots$. Par exemple dans le cas de deux paramètres, on considère dans l'espace $\mathbb{R}^3$ les points $M_i = (x_i, x_i', y_i)$ et il s'agit de trouver un plan affine P minimisant la somme des $d(M_i, P)^2$, où $d(M, P)$ est la distance du point M au plan P.

Si H est le projeté orthogonal de M sur P et si M' est le projeté de M dans la direction de l'axe des y, alors on a $MH = d(M,P) = MM'|\cos\alpha|$, où α est l'écart angulaire entre la normale à P et l'axe des y. Il revient donc au même de minimiser la somme des ${M_i'M_i}^2$.



Si l'équation du plan P est $y = ax + a'x' + b$, alors M_i' a pour coordonnées $(x_i, x_i', ax_i + a'x_i' + b)$ et $M_i'M_i = |y_i - ax_i - a'x_i' - b|$. On cherche donc des nombres a, a', b rendant minimum la somme des $(y_i - ax_i - a'x_i' - b)^2$.

Supposons que le nombre des données est $n \geqslant 3$ et notons

X, X' et Y les vecteurs de $\mathbb{R}^n$ de coordonnées (x_i), (x'_i) et (y_i). En appelant encore U le vecteur dont toutes les coordonnées sont égales à 1, la somme des $(y_i - ax_i - a'x'_i - b)^2$ est $\|Y - aX - a'X' - bU\|^2$: on projette donc orthogonalement Y sur le sous-espace $\mathscr{W}$ de $\mathbb{R}^n$ engendré par X, X', U ; en général, ces vecteurs sont indépendants et $\mathscr{W}$ est de dimension 3. Pour effectuer les calculs, on utilise la formule de la proposition précédente.

2. Matrices unitaires, matrices hermitiennes

Dans ce paragraphe, V est un espace hermitien de dimension n.

Notations. Si X est un vecteur-colonne de $\mathbb{K}^n$, on note $\overline{X}$ le vecteur obtenu en conjuguant tous les coefficients de X. De même, si A est une matrice, $\overline{A}$ est la matrice ayant pour coefficients les conjugués des coefficients de A.
Par définition de la transposée et du produit matriciel, on a

$$^t(\overline{A}) = \overline{^tA} \ , \quad \overline{AB} = \overline{A}\,\overline{B} \ \text{ et donc, si } A \text{ est inversible, } \overline{A^{-1}} = (\overline{A})^{-1}.$$

Expression du produit hermitien dans une base orthonormée

Si $e_1, e_2, \ldots, e_n$ est une base orthonormée de V et si u et u' sont des vecteurs de V, alors d'après la proposition page 206, on a $u \cdot u' = (^tU)\overline{U'}$, où U et U' sont les vecteurs-colonne des coordonnées de u et u' dans cette base.

2.1 Caractérisation matricielle d'une base orthonormée

Donnons-nous une base orthonormée $\mathscr{B}$ de V.
Soient $u_1, u_2, \ldots, u_n$ des vecteurs, $U_1, U_2, \ldots, U_n$ les vecteurs-colonne de leurs coordonnées dans la base $\mathscr{B}$ et U la matrice ayant $U_1, U_2, \ldots, U_n$ pour colonnes.
Puisque $\mathscr{B}$ est orthonormée, on a $u_i \cdot u_j = (^tU_i)\overline{U_j}$ et par définition du produit matriciel, ce nombre est le coefficient situé en i-ème ligne et j-ème colonne dans la matrice $(^tU)\overline{U}$.
Pour que les vecteurs $u_1, u_2, \ldots, u_n$ forment une base orthonormée de V, il faut et il suffit que l'on ait $(^tU_i)\overline{U_j} = 0$ si $i \neq j$ et $(^tU_i)\overline{U_i} = 1$ pour tous i et j, autrement dit que l'on ait $(^tU)\overline{U} = I_n$.

Définitions

Une matrice $U \in \mathscr{M}_n(\mathbb{C})$ telle que $(^tU)\overline{U} = I_n$ est dite *unitaire*. Une matrice $U \in \mathscr{M}_n(\mathbb{R})$ telle que $(^tU)U = I_n$ est dite *orthogonale*. On note $\mathbb{U}(n)$ l'ensemble des matrices unitaires de $\mathscr{M}_n(\mathbb{C})$ et $\mathbb{O}(n)$ l'ensemble des matrices orthogonales de $\mathscr{M}_n(\mathbb{R})$.

Proposition. *Soit $\mathscr{B}$ une base orthonormée de V. Pour que des vecteurs $u_1, u_2, \ldots, u_n$ de V forment une base orthonormée de V, il faut et il suffit que la matrice des coordonnées des u_i dans la base $\mathscr{B}$ soit unitaire si $\mathbb{K} = \mathbb{C}$, orthogonale si $\mathbb{K} = \mathbb{R}$.*

Une matrice de $\mathscr{M}_n(\mathbb{C})$ est unitaire si et seulement si ses colonnes forment une base orthonormée de $\mathbb{C}^n$; une matrice de $\mathscr{M}_n(\mathbb{R})$ est orthogonale si et seulement si ses colonnes forment une base orthonormée de $\mathbb{R}^n$.

Exemples

- Chacune des matrices $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, $\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$ et $\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$ est orthogonale.
- Si α et β sont des nombres complexes, les vecteurs $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ et $\begin{bmatrix} -\overline{\beta} \\ \overline{\alpha} \end{bmatrix}$ sont orthogonaux dans l'espace hermitien usuel $\mathbb{C}^2$. Il s'ensuit que si $|\alpha|^2 + |\beta|^2 = 1$, alors la matrice $\begin{bmatrix} \alpha & -\overline{\beta} \\ \beta & \overline{\alpha} \end{bmatrix}$ est unitaire.
- Une matrice est orthogonale si et seulement si elle est unitaire et à coefficients réels.

Propriétés des matrices unitaires ou orthogonales

i) *Toute matrice P unitaire (ou orthogonale) est inversible et P^{-1} est unitaire (ou orthogonale). On a $P^{-1} = {}^t\overline{P}$ si P est unitaire, $P^{-1} = {}^tP$ si P est orthogonale : pour inverser une matrice orthogonale, il suffit de la transposer.*

ii) *Le produit de deux matrices unitaires (ou orthogonales) est unitaire (ou orthogonale).*

iii) *Le déterminant d'une matrice unitaire est un nombre complexe de module 1.*
Le déterminant d'une matrice orthogonale est égal à ± 1.

Démonstration. Si P est une matrice carrée, on a $({}^tP)\overline{P} = I_n \Longleftrightarrow ({}^t\overline{P})P = \overline{I_n} = I_n$: une matrice P est donc unitaire si et seulement si $P^{-1} = {}^t\overline{P}$; posons alors $Q = P^{-1}$: puisque $Q = {}^t\overline{P}$, on a ${}^t\overline{Q} = P = Q^{-1}$, donc Q est unitaire et nous avons montré (i).
Si P et Q sont des matrices unitaires, alors ${}^t(PQ)\overline{PQ} = {}^tQ\,{}^tP\,\overline{P}\,\overline{Q} = {}^tQI_n\overline{Q} = {}^tQ\overline{Q} = I_n$, donc PQ est unitaire, d'où (ii).
Pour toute matrice carrée P à coefficients complexes, on a $\det\overline{P} = \overline{\det P}$, donc $\det\big(({}^tP)\overline{P}\big) = (\det{}^tP)(\det\overline{P}) = (\det P)(\overline{\det P}) = |\det P|^2$. Si P est une matrice unitaire, il vient $1 = \det I_n = \det\big(({}^tP)\overline{P}\big) = |\det P|^2$. Si de plus P est à coefficients réels, son déterminant est un nombre réel et l'égalité $|\det P|^2 = 1$ implique $\det P = \pm 1$. ■

2.2 Isométries d'un espace euclidien de dimension finie

Dans l'espace euclidien usuel de dimension 3, les rotations et les symétries sont des transformations qui conservent la distance : si f est une telle transformation, on a $\big\|f\big(\overrightarrow{AB}\big)\big\| = \|\overrightarrow{AB}\|$ pour tous points A et B, autrement dit f ne change pas la norme des vecteurs. Ces transformations seront étudiées à la fin du chapitre, mais nous présentons maintenant leurs propriétés générales.
Dans ce paragraphe, V est un espace euclidien de dimension n.

Définition

Une transformation linéaire $f : V \to V$ est une *isométrie de V* si $\|f(u)\| = \|u\|$ pour tout vecteur $u \in V$.

Propriétés d'une isométrie. *Si $f : V \to V$ est une isométrie, alors*

i) *pour tous vecteurs u et v de V, on a $f(u) \cdot f(v) = u \cdot v$ (conservation du produit scalaire) ;*

ii) *f est une bijection et f^{-1} est une isométrie.*

Démonstration. Pour tous vecteurs u et v, on a $2(u \cdot v) = \|u+v\|^2 - \|u\|^2 - \|v\|^2$. Si f est une isométrie, alors

$$\begin{aligned} 2\,f(u) \cdot f(v) &= \|f(u) + f(v)\|^2 - \|f(u)\|^2 - \|f(v)\|^2 = \|f(u+v)\|^2 - \|f(u)\|^2 - \|f(v)\|^2 \\ &= \|u+v\|^2 - \|u\|^2 - \|v\|^2 = 2\,u \cdot v\,. \end{aligned}$$

Si $f(u) = \mathbf{0}$, alors il vient $0 = \|f(u)\| = \|u\|$, donc $u = \mathbf{0}$: cela montre que le noyau de f est réduit au vecteur nul, donc la transformation linéaire f est bijective, d'après un corollaire page 173. ■

Caractérisation d'une isométrie. *Pour qu'une transformation linéaire $f : V \to V$ soit une isométrie, il faut et il suffit que la matrice de f dans une base orthonormée soit une matrice orthogonale.*

Démonstration. Soit A la matrice de f dans une base orthonormée $(e_1, e_2, \ldots, e_n)$ de V. Notons A_i la i-ème colonne de A : elle est formée des coordonnées de $f(e_i)$.
Supposons que f est une isométrie. Puisque e_i et e_j sont orthogonaux pour $i \neq j$, $f(e_i)$ et $f(e_j)$ le sont aussi, ce qui se traduit par $({}^tA_i)A_j = 0$; puisque e_i est de norme 1, $f(e_i)$ aussi, donc $({}^tA_i)A_i = 1$. Ces relations signifient que A est une matrice orthogonale. Réciproquement, si A est orthogonale, alors pour tout vecteur u de coordonnées X, on a $u \cdot u = ({}^tX)X$ et $f(u) \cdot f(u) = {}^t(AX)(AX) = ({}^tX)({}^tA)AX = ({}^tX)I_nX = ({}^tX)X$, d'où $\|f(u)\|^2 = \|u\|^2$. ■

Groupe des isométries de V. La composée de deux isométries est une isométrie, l'application identité est une isométrie et si f est une isométrie, alors f^{-1} aussi. Les isométries de V forment donc un groupe de transformations.

2.3 Diagonalisation des matrices hermitiennes

Définition
Une matrice $M \in \mathscr{M}_n(\mathbb{C})$ telle que ${}^tM = \overline{M}$ s'appelle une *matrice hermitienne*.

Une matrice à coefficients réels est hermitienne si et seulement si elle est symétrique.

Exemple. Toute matrice hermitienne de taille 2 est de la forme

$$\begin{bmatrix} x & a - b\boldsymbol{i} \\ a + b\boldsymbol{i} & y \end{bmatrix} = \frac{x+y}{2} I_2 + \frac{x-y}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + a \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{bmatrix},$$

où x, y, a, b sont des nombres réels. Les matrices hermitiennes $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ et $\begin{bmatrix} 0 & -\boldsymbol{i} \\ \boldsymbol{i} & 0 \end{bmatrix}$ s'appellent les *matrices de Pauli* (exercice 5, page 156).

Proposition. *Si $M \in \mathscr{M}_n(\mathbb{C})$ est une matrice hermitienne et si U est une matrice unitaire de taille n, alors $U^{-1}MU$ est hermitienne. Si $S \in \mathscr{M}_n(\mathbb{R})$ est symétrique et si P est une matrice orthogonale de taille n, alors $P^{-1}SP$ est symétrique.*

Démonstration. On a ${}^t(U^{-1}MU) = ({}^tU)({}^tM)({}^tU^{-1}) = \overline{U}^{-1}\,\overline{M}\,\overline{U}$ car ${}^t\overline{U} = U^{-1}$ et ${}^tM = \overline{M}$. Il vient ${}^t(U^{-1}MU) = \overline{U^{-1}MU}$, donc $U^{-1}MU$ est hermitienne. ■

Valeurs propres et vecteurs propres d'une matrice hermitienne

- *Les valeurs propres d'une matrice hermitienne sont des nombres réels. Des vecteurs propres associés à deux valeurs propres différentes sont orthogonaux dans $\mathbb{C}^n$.*
- *Pour une matrice symétrique réelle, des vecteurs propres associés à deux valeurs propres différentes sont orthogonaux dans $\mathbb{R}^n$.*

Démonstration. Soit M une matrice hermitienne de taille n et soient X, Y des vecteurs propres associés à des valeurs propres λ et μ. On a $MX = \lambda X$, $MY = \mu Y$ et donc

$$({}^tX)({}^tM)\overline{Y} = {}^t(MX)\overline{Y} = {}^t(\lambda X)\overline{Y} = \lambda{}^t(X\overline{Y}) = \lambda(X \cdot Y)$$

$$({}^tX)\overline{M}\overline{Y} = ({}^tX)\overline{MY} = ({}^tX)\overline{\mu Y} = ({}^tX)\bar{\mu}\overline{Y} = \bar{\mu}({}^tX)\overline{Y} = \bar{\mu}(X \cdot Y)\,.$$

Puisque ${}^tM = \overline{M}$, on en déduit l'égalité $\lambda(X \cdot Y) = \bar{\mu}(X \cdot Y)$, ou encore $(\lambda - \bar{\mu})X \cdot Y = 0$. Choisissons $X = Y$, donc $\lambda = \mu$: il vient $(\mu - \bar{\mu})\|X\|^2$, donc $\mu = \bar{\mu}$, car $\|X\|$ n'est pas nul. Supposons $\lambda \neq \mu$: alors $\lambda - \bar{\mu} = \lambda - \mu \neq 0$, donc $X \cdot Y = 0$. ■

Voici un résultat fondamental, ainsi que son corollaire.

Diagonalisation des matrices hermitiennes. *Si M est une matrice hermitienne de taille n, il existe une matrice unitaire U telle que $U^{-1}MU$ est diagonale à coefficients réels. En particulier, M est diagonalisable sur $\mathbb{C}$ et les colonnes de U forment une base orthonormée de $\mathbb{C}^n$ formée de vecteurs propres pour la transformation $X \mapsto MX$.*

Démonstration. Remarquons que le résultat est évident pour une matrice de taille 1 et montrons la première affirmation par récurrence sur la taille de la matrice. Soient λ une valeur propre de M et $e \in \mathbb{C}^n$ un vecteur propre pour λ, choisi de manière que $\|e\| = 1$. Soit W le sous-espace vectoriel de $\mathbb{C}^n$ orthogonal à e. Puisque W est de dimension $n-1$, il existe une base orthonormée $(w_2, \ldots, w_n)$ de W et les vecteurs $e, w_2, \ldots, w_n$ forment alors une base orthonormée de $\mathbb{C}^n$. Puisque $f(e) = \lambda e$, la transformation $f : X \mapsto MX$ a, dans cette base, une matrice de la forme $A = \begin{bmatrix} \lambda & * \\ 0 & M' \end{bmatrix}$, où M' est une matrice carrée de taille $n-1$. La matrice Q formée des vecteurs-colonne $e, w_2, \ldots, w_n$ est unitaire et d'après la formule du changement de base, on a $A = Q^{-1}MQ$. D'après la première proposition de ce paragraphe, A est une matrice hermitienne, ce qui implique à la fois que sa première ligne est $[\,\lambda\ 0 \cdots\ 0\,]$ et que M' est hermitienne. Par hypothèse de récurrence, il existe donc une matrice R' unitaire de taille $n-1$ telle que $R'^{-1}M'R'$ est diagonale. Posons $R = \begin{bmatrix} 1 & 0 \\ 0 & R' \end{bmatrix}$. La matrice R a ses vecteurs-colonne deux à deux orthogonaux et de norme 1, donc R est unitaire et $R^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & R'^{-1} \end{bmatrix}$. D'après les règles du produit matriciel, il vient

$$R^{-1}AR = \begin{bmatrix} 1 & 0 \\ 0 & R'^{-1} \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & M' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & R' \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & R'^{-1}M'R \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & D' \end{bmatrix}$$

où D' est diagonale. On a $R^{-1}AR = R^{-1}Q^{-1}MQR = (QR)^{-1}M(QR)$. Puisque les matrices Q et R sont unitaires, il en va de même de leur produit $U = QR$. En posant $D' = \mathrm{diag}(\lambda_2, \ldots, \lambda_n)$ et $\lambda_1 = \lambda$, on a donc $U^{-1}MU = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Puisque $U^{-1}MU$ est diagonale, les vecteurs-colonne de U sont vecteurs propres de la transformation $X \mapsto MX$: en effet, si U_i est le i-ème vecteur-colonne de U, alors $MU_i = MU\boldsymbol{E_i} = U\,\mathrm{diag}(\lambda_1, \cdots, \lambda_n)\boldsymbol{E_i} = U(\lambda_i\boldsymbol{E_i}) = \lambda_i U\boldsymbol{E_i} = \lambda_i U_i$. ■

Si S est une matrice symétrique à coefficients réels, elle est hermitienne, donc ses valeurs propres sont réelles ; puisque S est réelle, ses vecteurs propres le sont aussi, d'où la propriété suivante.

Diagonalisation des matrices symétriques réelles. *Si S est une matrice symétrique réelle de taille n, il existe une matrice orthogonale P telle que $P^{-1}SP$ est diagonale. En particulier, S est diagonalisable sur $\mathbb{R}$ et les colonnes de P forment une base orthonormée de $\mathbb{R}^n$ formée de vecteurs propres pour la transformation $X \mapsto SX$ de $\mathbb{R}^n$.*

Exemple. La matrice $S = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ est symétrique à coefficients réels, de polynôme caractéristique $\det(S - zI_3) = (1-z)(3-z)(4-z)$.

- Le vecteur $(1, 1, -2)$ est propre pour la valeur propre 1 et de norme $\sqrt{6}$,
- le vecteur $(1, -1, 0)$ est propre pour la valeur propre 3 et de norme $\sqrt{2}$,
- le vecteur $(1, 1, 1)$ est propre pour la valeur propre 4 et de norme $\sqrt{3}$.

Ces vecteurs sont deux à deux orthogonaux dans $\mathbb{R}^3$, donc les vecteurs

$$E_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \quad , \quad E_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \quad , \quad E_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

forment une base orthonormée de$\mathbb{R}^3$. En mettant E_1, E_2, E_3 en colonnes, on obtient la matrice orthogonale $P = \frac{1}{6} \begin{bmatrix} \sqrt{6} & 3\sqrt{2} & 2\sqrt{3} \\ \sqrt{6} & -3\sqrt{2} & 2\sqrt{3} \\ -2\sqrt{6} & 0 & 2\sqrt{3} \end{bmatrix}$ telle que $P^{-1}SP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}$.

Dans les applications, on considère souvent les produits scalaires mutuels de vecteurs $M_1, M_2, \ldots, M_p$ de $\mathbb{R}^n$ en formant la matrice dont le coefficient en position (i, j) est le produit scalaire $M_i \cdot M_j = ({}^tM_i)M_j$. En appelant $M = [\, M_1 \; \ldots \; M_p \,]$ la matrice de taille (n, p) ayant pour colonnes $M_1, \ldots, M_p$, on obtient ainsi la matrice carrée $[M_i{\cdot}M_j] = ({}^tM)M$, de taille p. Rappelons (page 210) que si $M_1, \ldots, M_p$ sont des vecteurs indépendants, la matrice $({}^tM)M$ est inversible.

Proposition. *Soit $M \in \mathscr{M}_{n,p}(\mathbb{R})$.*

- *La matrice $({}^tM)M$ est symétrique et ses valeurs propres sont des nombres réels positifs ou nuls.*
- *Supposons que M est une matrice carrée. Alors M est inversible si et seulement si les valeurs propres de $({}^tM)M$ sont toutes strictement positives.*

Démonstration. La matrice $A = ({}^tM)M$ est carrée de taille p et la transposée de A est $({}^tM)^t({}^tM) = ({}^tM)M = A$, donc A est symétrique. D'après le corollaire précédent, il existe une matrice orthogonale P telle que $P^{-1}AP = D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, où les λ_i son réels. Puisque $P^{-1} = {}^tP$, il vient $D = ({}^tP)AP = ({}^tP)({}^tM)MP = {}^t(MP)(MP)$ et pour tout vecteur-colonne $X \in \mathbb{R}^p$, on a $X \cdot DX = ({}^tX)DX = ({}^tX)^t(MP)MPX = {}^t(MPX)MPX = \|MPX\|^2 \geqslant 0$. En prenant pour X le vecteur canonique $\boldsymbol{E_i}$, on obtient $D\boldsymbol{E_i} = \lambda_i \boldsymbol{E_i}$ et $\lambda_i = \boldsymbol{E_i} \cdot D\boldsymbol{E_i} \geqslant 0$.

Supposons que M est une matrice carrée. On a $\det A = (\det {}^t M)(\det M) = (\det M)^2$. Puisque $\det A = \det D$ est le produit des λ_i, le déterminant de M est non nul si et seulement si les λ_i sont tous non nuls. ■

Construction de produits scalaires

Soit $S \in \mathscr{M}_n(\mathbb{R})$. Pour tous vecteurs-colonne X et X' appartenant à $\mathbb{R}^n$, le produit $({}^tX)SX'$ du vecteur-ligne ${}^tX$ par le vecteur-colonne SX' est une matrice de taille 1, c'est-à-dire un nombre réel ; une telle matrice est évidemment égale à sa transposée, donc $({}^tX)SX' = {}^t\big(({}^tX)SX'\big) = ({}^tX')({}^tS)X$. Si l'on suppose en outre que S est symétrique, alors il vient $({}^tX)SX' = ({}^tX')SX$.

Puisque le nombre $({}^tX)SX'$ dépend linéairement de X et de X', l'application $(X, X') \mapsto ({}^tX)SX'$ définira un produit scalaire sur $\mathbb{R}^n$ si l'on a la propriété de positivité : $({}^tX)SX > 0$ pour tout vecteur $X \neq 0$.

Définition

Une matrice S symétrique à coefficients réels est dite *définie positive* si $({}^tX)SX > 0$ pour tout vecteur-colonne $X \neq 0$.

Proposition. *Soit S une matrice symétrique à coefficients réels et de taille n.*

a) *Si r est la plus petite valeur propre de S et R la plus grande, alors pour tout vecteur* $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, *on a* $r\sum_{i=1}^n x_i^2 \leqslant ({}^tX)SX \leqslant R\sum_{i=1}^n x_i^2$.

b) *La matrice S est définie positive si et seulement si toutes ses valeurs propres sont strictement positives. Dans ce cas, $(X, X') \mapsto ({}^tX)SX'$ est un produit scalaire sur $\mathbb{R}^n$.*

Démonstration. D'après le corollaire précédent, il existe une matrice orthogonale P et une matrice diagonale $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ telle que $P^{-1}SP = D$. Soit $X \in \mathbb{R}^n$ un vecteur-colonne. Puisque P est inversible, il y a un vecteur Y tel que $X = PY$ et l'on a $({}^tX)SX = ({}^tY)({}^tP)SPY = ({}^tY)P^{-1}SPY$ car P est orthogonale. Il vient $({}^tX)SX = ({}^tY)DY = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$. Puisqu'on a $r \leqslant \lambda_i \leqslant R$ pour tout i, on en déduit $r\sum_{i=1}^n y_i^2 \leqslant ({}^tX)SX \leqslant R\sum_{i=1}^n y_i^2$. De plus, X et Y ont la même norme euclidienne, car la matrice P est orthogonale, donc $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2$. Si toutes les valeurs propres λ_i sont strictement positives, alors il vient $0 < r\|X\|^2 \leqslant ({}^tX)SX$ pour $X \neq 0$: ainsi la matrice S est définie positive et l'expression $({}^tX)SX'$ définit un produit scalaire sur $\mathbb{R}^n$. Réciproquement, s'il existe k tel que $\lambda_k \leqslant 0$, alors pour le vecteur canonique $Y = \boldsymbol{E_k}$, on a $({}^tY)DY = \lambda_k y_k^2 \leqslant 0$, donc $({}^tX)SX \leqslant 0$ si $X = P\boldsymbol{E_k}$. ■

Remarque

Dans la proposition, l'inégalité de gauche est une égalité si X est un vecteur propre pour la plus petite valeur propre de S ; celle de droite est une égalité si X est un vecteur propre pour la plus grande valeur propre.

Exemple 1. Cherchons à quelle condition la matrice symétrique réelle $S = \begin{bmatrix} p & q \\ q & r \end{bmatrix}$ est définie positive. Le polynôme caractéristique est $\begin{vmatrix} p-z & q \\ q & r-z \end{vmatrix} = z^2 - (p+r)z + pr - q^2$. Les racines sont positives si et seulement si leur produit $pr - q^2$ et leur somme $p + r$ sont positifs. Ces conditions sont équivalentes à $pr - q^2 > 0$ et $p > 0$, car $pr > q^2 \geqslant 0$ implique que p et r sont de même signe, donc du signe de leur somme. On retrouve ainsi les conditions qui ont permis, page 202, de construire un produit scalaire $xx' + q(xy' + x'y) + ryy' = [\,x\ \ y\,] \begin{bmatrix} p & q \\ q & r \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix}$ sur $\mathbb{R}^2$.

Exemple 2 : étude d'une ellipse. Soit $\mathscr{E}$ la courbe d'équation $5x^2 + 4xy + 2y^2 = 9$. On a

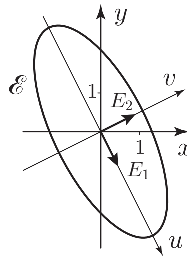
$$5x^2 + 4xy + 2y^2 = [\,x\ \ y\,] \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} .$$

La matrice $S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$ est symétrique, son polynôme caractéristique est $z^2 - 7z + 6 = (z-1)(z-6)$, ses valeurs propres sont strictement positives, donc S est définie positive. Le vecteur $(1, -2)$ est propre pour la valeur propre 1 et le vecteur orthogonal $(2, 1)$ est propre pour la valeur propre 6. Les vecteurs

$$E_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ -2 \end{bmatrix} \quad \text{et} \quad E_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

forment une base orthonormée de vecteurs propres de S. Si P est la matrice orthogonale de colonnes E_1, E_2, les coordonnées dans (E_1, E_2) d'un vecteur $X \in \mathbb{R}^2$ sont données par le vecteur-colonne U tel que $X = PU$. On a

$$\begin{aligned} 5x^2 + 4xy + 2y^2 &= {}^tXSX = {}^t(PU)SPU = ({}^tU)({}^tP)SPU , \quad \text{où } U = \begin{bmatrix} u \\ v \end{bmatrix}, \\ &= ({}^tU)P^{-1}SPU , \quad \text{car } {}^tP = P^{-1}, \\ &= ({}^tU) \begin{bmatrix} 1 & 0 \\ 0 & 6 \end{bmatrix} U = u^2 + 6v^2 . \end{aligned}$$



L'équation de $\mathscr{E}$ dans la base orthonormée (E_1, E_2) est donc simplement

$$u^2 + 6v^2 = 9 \qquad \textit{(équation réduite)}$$

Ainsi $\mathscr{E}$ est une ellipse d'axes dirigés par les vecteurs E_1 et E_2. Pour un point de l'ellipse, on a $-3 \leqslant u \leqslant 3$ et $-(1/2)\sqrt{6} \leqslant v \leqslant (1/2)\sqrt{6}$: le petit axe, dirigé par E_2, a pour longueur $\sqrt{6}$ et le grand axe a pour longueur 6.

Exemple 3 : un ellipsoïde. Reprenons la matrice symétrique $S = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ étudiée page 217. Soit K un nombre positif et $\mathscr{E}_K$ la surface d'équation

$$3x^2 + 3y^2 + 2z^2 + 2xz + 2yz = K^2 .$$

On a $\qquad 3x^2 + 3y^2 + 2z^2 + 2xz + 2yz = {}^tXSX$, pour tout $X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$.

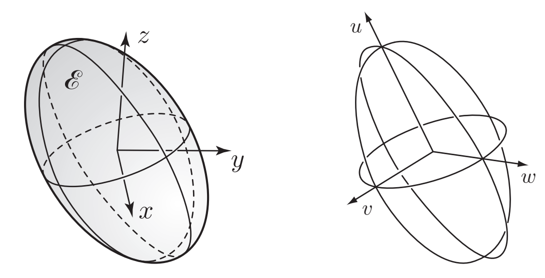
Pour étudier la surface, plaçons-nous dans la base orthonormée (E_1, E_2, E_3) formée des vecteurs propres de S. Les coordonnées de X dans cette base sont données par le vecteur-colonne U tel que $X = PU$, où P est la matrice de changement de base, c'est-à-dire la matrice de colonnes E_1, E_2, E_3. Il vient

$$({}^tX)SX = {}^t(PU)SPU = ({}^tU)({}^tP)SPU = ({}^tU)P^{-1}SPU \ , \quad \text{car } {}^tP = P^{-1}$$
$$= ({}^tU)\begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} U = u^2 + 3v^2 + 4w^2 \ , \quad \text{où } U = \begin{bmatrix} u \\ v \\ w \end{bmatrix}.$$

Dans la base orthonormée (E_1, E_2, E_3), l'équation de $\mathscr{E}_K$ est donc simplement

$$u^2 + 3v^2 + 4w^2 = K^2 \qquad \textit{(équation réduite)}$$

Si l'on coupe la surface par le plan des coordonnées u, v (en faisant $w = 0$), on obtient l'ellipse d'équation $u^2 + 3v^2 = K^2$: elle est centrée à l'origine et ses axes sont les axes des coordonnées u et v. De même, l'intersection de la surface avec les autres plans de coordonnées sont des ellipses ayant pour axes les directions des coordonnées u, v ou w. Ainsi la surface est un ellipsoïde d'axes dirigés par E_1, E_2, E_3. Pour les points de $\mathscr{E}_K$ qui sont sur les axes, on a $-K \leqslant u \leqslant K$, $-K/\sqrt{3} \leqslant v \leqslant K/\sqrt{3}$ et $-K/2 \leqslant w \leqslant K/2$: le plus grand axe de l'ellipsoïde, porté par u, a pour longueur $2K$; l'axe porté par v a pour longueur $2K/\sqrt{3}$ et le plus petit axe, porté par w, a pour longueur K.



On doit parfois comparer deux produits scalaires : voici comment calculer la valeur extrême que peut prendre leur rapport.

Proposition. *Soient A et B des matrices de même taille n, symétriques définies positives. Pour tout vecteur-colonne non nul $X \in \mathbb{R}^n$, on a $\dfrac{({}^tX)AX}{({}^tX)BX} \leqslant \lambda$, où λ est la plus grande valeur propre de $B^{-1}A$.*

Démonstration. Il existe une matrice inversible C telle que $B = ({}^tC)C$ (exercice 5 page 238). Pour tout vecteur $X \in \mathbb{R}^n$, on a donc $({}^tX)BX = ({}^tX)({}^tC)CX = ({}^tY)Y$, où $Y = CX$. Puisque $X = C^{-1}Y$, il vient, si $Y \neq 0$,

$$\frac{({}^tX)AX}{({}^tX)BX} = \frac{({}^tY)({}^tC^{-1})AC^{-1}Y}{({}^tY)Y} = \frac{({}^tY)SY}{({}^tY)Y}, \quad \text{où } S = ({}^tC^{-1})AC^{-1}.$$

La matrice S est symétrique. On a $CB^{-1} = C(C^{-1}{}^tC^{-1}) = {}^tC^{-1}$, donc $CB^{-1}A = {}^tC^{-1}A$ et $S = C(B^{-1}A)C^{-1}$. Les matrices $B^{-1}A$ et S ont même polynôme caractéristique (page 171), donc λ est aussi la plus grande valeur propre de S. D'après la proposition précédente, on a $\frac{({}^tY)SY}{({}^tY)Y} \leqslant \lambda$ pour tout $Y \neq 0$ et donc $\frac{({}^tX)AX}{({}^tX)BX} \leqslant \lambda$ pour tout $X \neq 0$. ■

3. Géométrie euclidienne

Soit E un espace euclidien de dimension 3.

Si M et M' sont des points de E, leur distance est par définition $\| \overrightarrow{MM'} \|$.

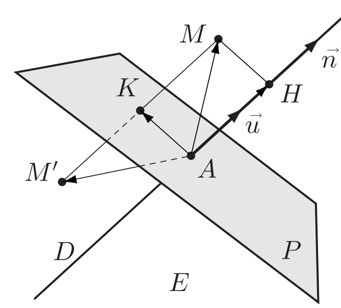
3.1 Projections et symétries

Considérons dans E une droite affine D passant par un point A et dirigée par un vecteur non nul $\vec{n}$. Soit P le plan orthogonal à $\vec{n}$ passant par A.

Puisque le vecteur $\vec{u} = \frac{1}{\| \vec{n} \|}\vec{n}$ est de norme 1, la projection orthogonale d'un vecteur $\vec{v}$ sur la droite vectorielle engendrée par $\vec{n}$ est $p(\vec{v}) = (\vec{v} \cdot \vec{u})\vec{u} = \frac{1}{\|\vec{n}\|^2}(\vec{v} \cdot \vec{n})\vec{n}$, d'après le théorème de la projection page 209. La projection de $\vec{v}$ sur le plan vectoriel orthogonal à $\vec{n}$ est $\vec{v} - p(\vec{v})$, d'où le résultat suivant :

Pour tout point $M \in E$,

- le projeté de M sur D est le point H défini par $\overrightarrow{AH} = \frac{\overrightarrow{AM} \cdot \vec{n}}{\| \vec{n} \|^2}\vec{n}$,
- le projeté de M sur P est le point K défini par $\overrightarrow{AK} = \overrightarrow{AM} - \overrightarrow{AH}$.
- le symétrique de M par rapport à P est le point M' tel que $\overrightarrow{AM'} = \overrightarrow{AM} - 2\overrightarrow{AH}$.



Distance à un plan. *Soit P un plan affine de E.*

- *Si P est le plan passant par un point A et orthogonal au vecteur $\vec{n} \neq \mathbf{0}$, la distance de M à P est* $\frac{| \overrightarrow{AM} \cdot \vec{n} |}{\| \vec{n} \|}$.
- *Si le plan P a pour équation $ax+by+cz+d=0$ dans un repère orthonormé de E et si M a pour coordonnées (x_M, y_M, z_M), la distance de M à P est* $\frac{|ax_M + by_M + cz_M + d|}{\sqrt{a^2+b^2+c^2}}$.

Démonstration. Avec les notations précédentes, la distance de M à P est $AH = \| \overrightarrow{AH} \| = \frac{|\overrightarrow{AM} \cdot \vec{n}|}{\|\vec{n}\|^2}\| \vec{n} \| = \frac{|\overrightarrow{AM} \cdot \vec{n}|}{\|\vec{n}\|}$, d'où la première formule. Si P a pour équation $ax+by+cz+d=0$ dans un repère orthonormé de E, alors le vecteur $\vec{n}$ de coordonnées (a, b, c) est orthogonal à P et si le point $A \in P$ a pour coordonnées (x_A, y_A, z_A), on a $\overrightarrow{AM} \cdot \vec{n} =$

$(x_M - x_A)a + (y_M - y_A)b + (z_M - z_A)c = x_M a + y_M b + z_M c + d$, car $d = -ax_A - bx_A - cx_A$.

La distance de M à P est donc $\frac{|\overrightarrow{AM}\cdot\vec{n}|}{\|\vec{n}\|} = \frac{|ax_M + by_M + cz_M c + d|}{\sqrt{a^2+b^2+c^2}}$. ■

Symétrie par rapport à un plan. Donnons-nous un plan P par son équation $ax+by+cz+d=0$ dans un repère orthonormé de E, où nous supposons $a^2+b^2+c^2=1$ (on dit que l'équation est « normale »). Le vecteur $\vec{n}$ de coordonnées (a,b,c) est orthogonal à P et de norme 1. Soit M un point de coordonnées (x,y,z). Si A est un point de P, le symétrique M' de M est défini par

$$(*) \qquad \overrightarrow{AM'} = \overrightarrow{AM} - 2\overrightarrow{AH} = \overrightarrow{AM} - 2(\overrightarrow{AM}\cdot\vec{n})\vec{n}$$

On a aussi (voir la figure précédente) $\overrightarrow{AM'} = \overrightarrow{AK} - \overrightarrow{AH}$ et $\overrightarrow{AM} = \overrightarrow{AK} + \overrightarrow{AH}$, donc $\| \overrightarrow{AM'} \| = \| \overrightarrow{AM} \|$, car les vecteurs $\overrightarrow{AK}$ et $\overrightarrow{AH}$ sont orthogonaux.

Si (p,q,r) sont les coordonnées de $\overrightarrow{AM}$, alors d'après $(*)$, celles de $\overrightarrow{AM'}$ sont

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} - 2(ap+bq+cr)\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} (1-2a^2)p - 2abq - 2acr \\ -2abp + (1-2b^2)q - 2bcr \\ -2acp - 2bcq + (1-2c^2)r \end{bmatrix} = \begin{bmatrix} 1-2a^2 & -2ab & -2ac \\ -2ab & 1-2b^2 & -2bc \\ -2ac & -2bc & 1-2c^2 \end{bmatrix}\begin{bmatrix} p \\ q \\ r \end{bmatrix}$$

Prenons pour A le point de coordonnées $(-da, -db, -dc)$ qui est bien dans P, puisque ses coordonnées en vérifient l'équation. Les coordonnées (x,y,z) et (x',y',z') de M et M' sont reliées par la relation

$$\begin{bmatrix} x'+da \\ y'+db \\ z'+dc \end{bmatrix} = \begin{bmatrix} 1-2a^2 & -2ab & -2ac \\ -2ab & 1-2b^2 & -2bc \\ -2ac & -2bc & 1-2c^2 \end{bmatrix}\begin{bmatrix} x+da \\ y+db \\ z+dc \end{bmatrix} = S\begin{bmatrix} x+da \\ y+db \\ z+dc \end{bmatrix}$$

Cas d'un plan passant par l'origine. On a alors $d=0$ et P est un plan vectoriel. La symétrie orthogonale par rapport à P est une transformation linéaire s de E. Le calcul ci-dessus montre que sa matrice S dans une base orthonormée est symétrique. En appelant O l'origine, on a $s(\overrightarrow{OM}) = \overrightarrow{OM'}$ et $\| \overrightarrow{OM'} \| = \| \overrightarrow{OM} \|$: s est donc une isométrie et la matrice S est aussi orthogonale.

On a $s(\vec{n}) = -\vec{n}$ et pour tout vecteur $\vec{v} \in P$, on a $s(\vec{v}) = \vec{v}$: les vecteurs propres de s pour la valeur propre 1 sont les vecteurs de P et le vecteur $\vec{n}$ est propre pour la valeur propre -1. Si l'on choisit une base quelconque $(\vec{e},\vec{f})$ de P , la matrice de s dans la base $(\vec{e},\vec{f},\vec{n})$ est donc $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$. Sur cette matrice, on voit que le déterminant de s est -1 : toute matrice d'une symétrie par rapport à un plan est donc aussi de déterminant -1.

3.2 Produit vectoriel

Dans ce paragraphe, on se donne une base orthonormée $\mathscr{B} = (e_1, e_2, e_3)$ de E.

Bases orthonormées directes

Soient u_1,u_2,u_3 des vecteurs de E et U_1,U_2,U_3 leurs vecteurs-colonne de coordonnées dans la base $\mathscr{B}$.

Les vecteurs u_1,u_2,u_3 forment une base orthonormée si et seulement si la matrice $U = [\,U_1 \;\; U_2 \;\; U_3\,]$ est orthogonale. Comme une matrice orthogonale a pour déterminant 1 ou -1, il y a deux sortes de bases : celles pour lesquelles $\det U = 1$ et celles où $\det U = -1$.

> **Définition**
> Si U est une matrice orthogonale de déterminant 1, on dit que la base orthonormée (u_1, u_2, u_3) de E est *directe* (relativement à la base $\mathscr{B}$).

Si l'on se donne deux vecteurs u_1 et u_2 orthogonaux et de norme 1, il y a deux vecteurs u_3 et $u_3' = -u_3$ de norme 1 et orthogonaux à u_1 et u_2 : en effet, les vecteurs orthogonaux au plan engendré par u_1 et u_2 forment une droite. Puisqu'on a $\det(U_1, U_2, -U_3) = -\det(U_1, U_2, U_3)$, l'une des bases orthonormée (u_1, u_2, u_3) ou $(u_1, u_2, -u_3)$ est directe et l'autre ne l'est pas.

Dans l'espace euclidien usuel, on dispose de la base canonique comme base $\mathscr{B}$ de référence. Voici un moyen mécanique pour caractériser les bases orthonormées directes (u_1, u_2, u_3) : *le vecteur u_3 a le sens de déplacement d'un tire-bouchon posé sur l'axe dirigé par u_3 et que l'on tourne en allant de u_1 vers u_2.*

Dans la figure, prenons (e_1, e_2, e_3) comme base de référence : alors (u_1, u_2, u_3) est directe et les bases (u_2, u_1, u_3) et (v_1, v_2, v_3) sont indirectes.

Définition du produit vectoriel

Rappelons que (e_1, e_2, e_3) est une base orthonormée de E.

Soient $u = pe_1 + qe_2 + re_3$, $u' = p'e_1 + q'e_2 + r'e_3$ et U, U' les vecteurs-colonne de $\mathbb{R}^3$ correspondants. Pour tout vecteur $w = xe_1 + ye_2 + ze_3$, de vecteur-colonne W, on a

$$\det(U, U', W) = \begin{vmatrix} p & p' & x \\ q & q' & y \\ r & r' & z \end{vmatrix} = \begin{vmatrix} q & q' \\ r & r' \end{vmatrix} x - \begin{vmatrix} p & p' \\ r & r' \end{vmatrix} y + \begin{vmatrix} p & p' \\ q & q' \end{vmatrix} z\,.$$

> **Définition**
> Le vecteur $\begin{vmatrix} q & q' \\ r & r' \end{vmatrix} e_1 - \begin{vmatrix} p & p' \\ r & r' \end{vmatrix} e_2 + \begin{vmatrix} p & p' \\ q & q' \end{vmatrix} e_3$ s'appelle le *produit vectoriel de u et u'* et se note $u \wedge u'$.

Pour les vecteurs de la base $\mathscr{B}$, on a ainsi : $e_1 \wedge e_2 = e_3$, $e_2 \wedge e_3 = e_1$ et $e_3 \wedge e_1 = e_2$.
Puisque la base $\mathscr{B}$ est orthonormée, l'expression ci-dessus de $\det(U, U', W)$ est le produit scalaire des vecteurs $u \wedge u'$ et w.

> $(u \wedge u') \cdot w = \det(U, U', W)$ *pour tous vecteurs u, v, w de coordonnées U, V, W dans une base orthonormée.*

Propriétés du produit vectoriel. *Soient u et u' des vecteurs de E.*

i) *Le produit vectoriel est antisymétrique et linéaire par rapport à chaque variable :*

$$u \wedge u' = -(u' \wedge u)$$
$$(u_1 + u_2) \wedge u' = u_1 \wedge u' + u_2 \wedge u' \quad \text{et} \quad (\lambda u) \wedge u' = \lambda(u \wedge u')$$
$$u \wedge (u'_1 + u'_2) = u \wedge u'_1 + u \wedge u'_2 \quad \text{et} \ u \wedge (\lambda u'_1) = \lambda(u \wedge u'_1).$$

ii) *Le vecteur $u \wedge u'$ est orthogonal à u et à u'.*

iii) *$u \wedge u'$ est nul si et seulement si les vecteurs u et u' sont colinéaires.*

iv) *Dans toute base orthonormée directe, les coordonnées de $u \wedge u'$ sont données par la formule de la définition.*

v) *Si u et u' sont orthogonaux et de norme 1, alors $(u, u', u \wedge u')$ est une base orthonormée directe de E.*

Démonstration. Les propriétés (i) se voient sur la définition du produit vectoriel en utilisant la linéarité et l'antisymétrie des déterminants. On a $(u \wedge u') \cdot u = \det(U, U', U) = 0$, donc $u \wedge u'$ est orthogonal à u ; de même, $u \wedge u'$ est orthogonal à u'. En appliquant la formule $u \wedge u' = -u' \wedge u$ à $u = u'$, on obtient $u \wedge u = \mathbf{0}$. On en déduit que si u et u' sont colinéaires, alors $u \wedge u'$ est nul. Si u et u' sont indépendants, on peut trouver un vecteur w tel que u, u', w soit une base de E (théorème de la base incomplète) ; alors $(u \wedge u') \cdot w = \det(U, U', W)$ n'est pas nul, donc $u \wedge u'$ n'est pas le vecteur nul.
Soit $\mathscr{B}_1$ une base orthonormée directe de E. La matrice de changement de base de $\mathscr{B}$ vers $\mathscr{B}_1$ est une matrice P orthogonale. Avec les notations précédentes, les coordonnées de u, u', w dans la base $\mathscr{B}_1$ sont les vecteurs-colonne U_1, U'_1, W_1 tels que $U = PU_1$, $U' = PU'_1$, $W = PW_1$. On a l'égalité matricielle $[U \ U' \ W] = [PU_1 \ PU'_1 \ PW_1] = P[U_1 \ U'_1 \ W_1]$. Comme le déterminant d'un produit de matrices est le produit des déterminants, on en déduit $\det(U, U', W) = (\det P)\det(U_1, U'_1, W_1) = \det(U_1, U'_1, W_1)$, car la base $\mathscr{B}_1$ étant directe, le déterminant de P vaut 1. Si l'on prend pour w les vecteurs de $\mathscr{B}$, ces déterminants sont les coordonnées de $u \wedge u'$ dans $\mathscr{B}$ et $\mathscr{B}_1$ respectivement : les coordonnées du vecteur $u \wedge u'$ sont donc données par les mêmes formules, que l'on calcule dans la base $\mathscr{B}$ ou dans la base $\mathscr{B}_1$.
Il reste à montrer la dernière affirmation. Supposons u et u' orthogonaux et de norme 1. Soit w le vecteur tel que $\mathscr{B}' = (u, u', w)$ est une base orthonormée directe. Dans cette base, les coordonnées de u sont $(1, 0, 0)$, celles de u' sont $(0, 1, 0)$, donc celles $u \wedge u'$ sont $(0, 0, 1)$: c'est donc que l'on a $u \wedge u' = w$. ■

Produit mixte. Un déterminant d'ordre 3 ne change pas si l'on permute circulairement ses colonnes : en effet, une permutation circulaire des trois colonnes se compose de deux échanges et pour chaque échange, le déterminant change de signe. Puisque le produit scalaire $(u_1 \wedge u_2) \cdot u_3$ est le déterminant des coordonnées de u_1, u_2, u_3 dans une base orthonormée, on en déduit les égalités

$$(u_1 \wedge u_2) \cdot u_3 = (u_3 \wedge u_1) \cdot u_2 = (u_2 \wedge u_3) \cdot u_1$$

Le nombre $(u_1 \wedge u_2) \cdot u_3$ s'appelle le *produit mixte* des vecteurs u_1, u_2, u_3.

Formulaire

(1) $\|u \wedge v\|^2 + (u \cdot v)^2 = \|u\|^2 \, \|v\|^2$.

(2) $u \wedge (v \wedge w) = (u \cdot w)v - (u \cdot v)w$ *(formule du double produit vectoriel)*

Démonstration de (1). Si u et v sont colinéaires, alors $u \wedge v$ est nul et la première formule est le cas d'égalité dans l'inégalité de Cauchy-Schwarz. Supposons u et v orthogonaux et non nuls et posons $u = \|u\| f_1$, $v = \|v\| f_2$. Puisque f_1 et f_2 sont orthogonaux et de norme 1, on peut choisir un vecteur f_3 tel que $\mathscr{B}' = (f_1, f_2, f_3)$ soit une base orthonormée directe de E. Les coordonnées de u et v dans cette base sont $(\|u\|, 0, 0)$ et $(0, \|v\|, 0)$, donc $u \wedge v = \begin{vmatrix} \|u\| & 0 \\ 0 & \|v\| \end{vmatrix} f_3 = \|u\| \, \|v\| f_3$. Le produit scalaire $u \cdot v$ étant nul, la formule (1) est vraie dans ce cas. Dans le cas général de deux vecteurs u et v indépendants, il y a un nombre réel λ tel que $v' = v - \lambda u$ soit orthogonal à u (procédé de Gram-Schmidt, page 206). On a $u \wedge v' = u \wedge v - u \wedge (\lambda u) = u \wedge v$, $\|v\|^2 = \|v' + \lambda u\|^2 = \|v'\|^2 + \lambda^2 \|u\|^2$ et $u \cdot v = u \cdot v' + \lambda \|u\|^2 = \lambda \|u\|^2$, donc

$$\begin{aligned} \|u \wedge v\|^2 = \|u \wedge v'\|^2 &= \|u\|^2 \, \|v'\|^2 \ , \quad \text{car } v' \text{ est orthogonal à } u \\ &= \|u\|^2 \left(\|v\|^2 - \lambda^2 \|u\|^2\right) = \|u\|^2 \, \|v\|^2 - (u \cdot v)^2 \, . \end{aligned}$$

■

La première formule permet de calculer la norme d'un produit vectoriel. Si u et v sont des vecteurs non nuls, alors d'après cette formule, il existe un unique nombre réel $\theta \in [0, \pi]$ tel que $\|u \wedge v\| = \|u\| \, \|v\| \sin\theta$ et $u \cdot v = \|u\| \, \|v\| \cos\theta$.

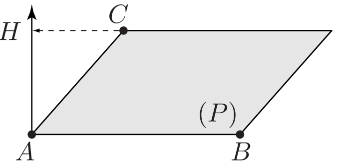
Définition

Le nombre $\theta \in [0, \pi]$ tel que $\|u \wedge v\| = \|u\| \, \|v\| \sin\theta$ et $u \cdot v = \|u\| \, \|v\| \cos\theta$ s'appelle *l'écart angulaire* des vecteurs u et v.

Exemples d'utilisation du produit vectoriel

Aire d'un parallélogramme

Soient A, B, C trois points non alignés d'un plan euclidien inclus dans E. Construisons le parallélogramme (P) sur ces trois points.

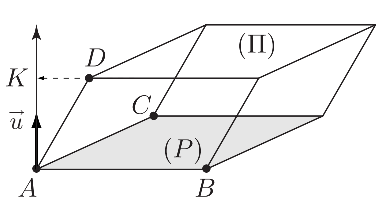


Projetons C en H sur la droite orthogonale à $\overrightarrow{AB}$ menée par A. L'aire de (P) est le produit des longueurs $AB \, AH$.
Puisque les vecteurs $\overrightarrow{AB}$ et $\overrightarrow{AH}$ sont orthogonaux, on a $\| \overrightarrow{AB} \| \, \| \overrightarrow{AH} \| = \| \overrightarrow{AB} \wedge \overrightarrow{AH} \|$.
Or $\overrightarrow{AB} \wedge \overrightarrow{AH} = \overrightarrow{AB} \wedge \overrightarrow{AC} + \overrightarrow{AB} \wedge \overrightarrow{CH}$ et $\overrightarrow{AB} \wedge \overrightarrow{CH} = \mathbf{0}$ car $\overrightarrow{CH}$ est colinéaire à $\overrightarrow{AB}$.
On en déduit que

l'aire du parallélogramme (P) est $\| \overrightarrow{AB} \wedge \overrightarrow{AC} \|$.

Volume d'un parallélépipède

Soit (Π) le parallélépipède construit avec le parallélogramme (P) et le point D de l'espace.



Projetons D en K sur la droite passant par A et orthogonale au plan de (P) ; choisissons un vecteur directeur $\vec{u}$ de cette droite tel que $\| \vec{u} \| = 1$.
Les vecteurs $\overrightarrow{AB} \wedge \overrightarrow{AC}$ et $\vec{u}$, tous deux orthogonaux au plan ABC, sont colinéaires : on a $\overrightarrow{AB} \wedge \overrightarrow{AC} = \lambda \vec{u}$, donc $|\lambda| = \| \overrightarrow{AB} \wedge \overrightarrow{AC} \|$ est l'aire du parallélogramme (P). Pour avoir

le volume du parallélépipède, il faut multiplier l'aire du parallélogramme par la hauteur AK. Puisque $AK = \left| \overrightarrow{AD} \cdot \vec{u} \right|$, ce volume est $|\lambda|\, AK = \left|\lambda \vec{u} \cdot \overrightarrow{AD}\right| = \left|(\overrightarrow{AB} \wedge \overrightarrow{AC}) \cdot \overrightarrow{AD}\right|$.

Si $U_{\overrightarrow{AB}}$, $U_{\overrightarrow{AC}}$ et $U_{\overrightarrow{AD}}$ sont les vecteurs-colonne des coordonnées de $\overrightarrow{AB}$, $\overrightarrow{AC}$, $\overrightarrow{AD}$ dans une base orthonormée, on a donc

$$\text{volume de } (\Pi) = \left|(\overrightarrow{AB} \wedge \overrightarrow{AC}) \cdot \overrightarrow{AD}\right| = \left|\det(U_{\overrightarrow{AB}}, U_{\overrightarrow{AC}}, U_{\overrightarrow{AD}})\right|.$$

On en déduit une autre expression de la distance d'un point D à un plan :

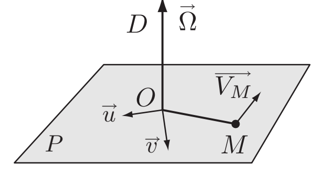
$$\text{distance de } D \text{ au plan } (ABC) = \frac{\left|\det(U_{\overrightarrow{AB}}, U_{\overrightarrow{AC}}, U_{\overrightarrow{AD}})\right|}{\| \overrightarrow{AB} \wedge \overrightarrow{AC} \|}.$$

Mouvement de rotation autour d'un axe

Si $\vec{w} \in E$ est un vecteur de norme 1, on peut trouver des vecteurs $\vec{u}$ et $\vec{v}$ tels que $(\vec{u}, \vec{v}, \vec{w})$ est une base orthonormée directe de E.

En effet, si P est le plan orthogonal à $\vec{w}$, alors pour toute base orthonormée $(\vec{u}, \vec{v})$ de P, $\vec{u} \wedge \vec{v}$ est de norme 1 et colinéaire à $\vec{w}$, donc on a $\vec{u} \wedge \vec{v} = \pm \vec{w}$. Puisque $\vec{u} \wedge \vec{v} = -\vec{v} \wedge \vec{u}$, il suffit de mettre les vecteurs dans le bon ordre pour obtenir une base orthonormée $(\vec{u}, \vec{v})$ de P telle que $\vec{u} \wedge \vec{v} = \vec{w}$.

Soit D la droite vectorielle engendrée par $\vec{w}$. La donnée du couple $(\vec{u}, \vec{v})$ définit un sens de rotation autour de D : celui qui amène $\vec{u}$ sur $\vec{v}$.



Pour repérer le sens d'un mouvement de rotation autour de D, il suffit donc de se donner le vecteur $\vec{w}$.

Le *vecteur rotation* du mouvement est par définition le vecteur $\vec{\Omega} = \omega \vec{w}$, où $\omega > 0$ est la vitesse angulaire ; le vecteur $\vec{\Omega}$ contient toutes les informations sur le mouvement : son sens indique à la fois l'axe et le sens de rotation, sa norme ω donne la vitesse de rotation.

Soit M un point de E en rotation autour de D et soit O un point de D.

- Le vecteur vitesse de M est $\overrightarrow{V_M} = \vec{\Omega} \wedge \overrightarrow{OM}$;
- si M est un point pesant de masse m, le *moment cinétique* de M en O est le vecteur $\overrightarrow{\sigma_O} = \overrightarrow{OM} \wedge m\overrightarrow{V_M}$.

Choisissons un repère orthonormé $(O; \vec{i}, \vec{j}, \vec{k})$ d'origine O, appelons (p, q, r) les coordonnées de $\vec{\Omega}$ et (x, y, z) celles de $\overrightarrow{OM}$. Les coordonnées du vecteur $\overrightarrow{V_M}$ sont $v_x = qz - ry$, $v_y = rx - pz$, $v_z = py - qx$ et les coordonnées de $\overrightarrow{\sigma_O}$ sont

$$\begin{bmatrix} m(yv_z - zv_y) \\ m(zv_x - xv_z) \\ m(xv_y - yv_x) \end{bmatrix} = \begin{bmatrix} m(y^2 + z^2)p - (mxy)q - (mxz)r \\ -(mxy)p + m(x^2 + z^2)q - (myz)r \\ -(mxz)p - (myz)q + m(x^2 + y^2)r \end{bmatrix}$$

$$= \begin{bmatrix} m(y^2 + z^2) & -mxy & -mxz \\ -mxy & m(x^2 + z^2) & -myz \\ -mxz & -myz & m(x^2 + y^2) \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix}$$

Pour un solide (S) en rotation autour de l'axe fixe D, les coordonnées du moment cinétique au point O (dans un repère d'origine O) sont donc données par

$$\begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} = J(O) \begin{bmatrix} p \\ q \\ r \end{bmatrix}, \text{ où}$$

$$I_{xx} = \iiint_S (y^2 + z^2)\, dm \quad , \quad I_{yy} = \iiint_S (x^2 + z^2)\, dm \quad , \quad I_{zz} = \iiint_S (x^2 + y^2)\, dm$$

$$I_{xy} = I_{yx} = \iiint_S xy\, dm \quad , \quad I_{xz} = I_{zx} = \iiint_S xz\, dm \quad , \quad I_{yz} = I_{zy} = \iiint_S yz\, dm$$

dm étant l'élément de masse infinitésimale placé au point (x, y, z).

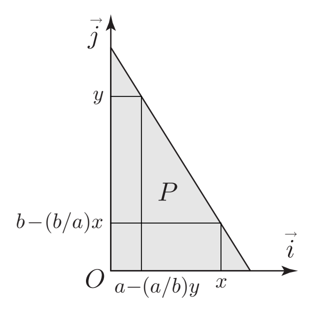
La matrice $J(O)$ est la *matrice d'inertie de* (S) *au point* O : elle ne dépend que du solide (S) et du point O. Cette matrice permet de calculer le moment cinétique en O d'un mouvement de rotation de autour de n'importe quel axe passant par O.

D'après l'égalité ci-dessus, le moment cinétique n'est dans la direction de l'axe que si le vecteur rotation est un vecteur propre de la matrice d'inertie. Puisque la matrice $J(O)$ est symétrique, il existe une base orthonormée de vecteurs propres : un axe ayant pour direction l'un de ces vecteurs s'appelle un *axe principal d'inertie* du solide au point O.

Les composantes du moment cinétique qui sont transverses à l'axe de rotation sont nuisibles aux grandes vitesses : c'est pourquoi on essaye d'équilibrer les pièces en rotation rapide pour que l'axe de rotation soit un axe principal d'inertie.

Pour étudier un mouvement de rotation, on a besoin d'un principe de mécanique, le « théorème du moment cinétique », valable quand le repère et le point O sont fixes : *la dérivée* $\dfrac{d\vec{\sigma}_O}{dt}$ *du moment cinétique par rapport au temps est égal à* $\mathscr{M}_O = \int_S \left(\overrightarrow{OM} \wedge \vec{F}\right)$, *moment en* O *des forces extérieures* $\vec{F}$ *appliquées au solide.*

Exemple. Considérons une plaque homogène P en forme de triangle rectangle en O. Choisissons un repère orthonormé $(O; \vec{i}, \vec{j}, \vec{k})$ d'origine O, les vecteurs $\vec{i}$ et $\vec{j}$ étant portés par les petits côtés de la plaque ; le vecteur $\vec{k}$ est orthogonal au plan de la plaque. Calculons la matrice d'inertie de la plaque au point O.



En tout point de la plaque, la coordonnée z est nulle, donc $I_{xz} = I_{yz} = 0$ et l'on a simplement $I_{xx} = \iint_P y^2\, dm$, $I_{yy} = \iint_P x^2\, dm$ et $I_{zz} = I_{yy} + I_{xx}$.

Notons a et b les longueurs des petits côtés de la plaque et m sa masse. Pour y fixé entre 0 et b, l'abscisse x d'un point de la plaque varie entre 0 et $a - (a/b)y$, donc

$$I_{xx} = m \int_0^b y^2 \left(\int_0^{a-(a/b)y} dx \right) dy = m \int_0^b y^2 \left(a - \frac{ay}{b} \right) dy = m \left(a\frac{b^3}{3} - \frac{a}{b}\frac{b^4}{4} \right) = \frac{mab^3}{12}\,.$$

Pour calculer I_{yy}, il suffit d'échanger les rôles de a et b : $I_{yy} = \dfrac{ma^3b}{12}$. Enfin, on a

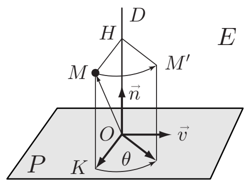
$$I_{xy} = m\int_0^b y\left(\int_0^{a-(a/b)y} x\,dx\right)dy = m\int_0^b y\frac{1}{2}\left(a-\frac{ay}{b}\right)^2 dy = \frac{ma^2b^2}{24}\ ,$$

et $I_{zz} = I_{yy} + I_{xx} = \dfrac{mab(a^2+b^2)}{12}$.

La matrice d'inertie en O est donc $J(O) = \dfrac{mab}{24}\begin{bmatrix} 2b^2 & -ab & 0 \\ -ab & 2a^2 & 0 \\ 0 & 0 & 2a^2+2b^2 \end{bmatrix}$. Le vecteur $\vec{k}$ est propre, ainsi que les vecteurs du plan xOy de pente $(1/\tau)\left(\tau^2 - 1 \pm \sqrt{1-\tau^2+\tau^4}\right)$, où $\tau = b/a$: ce sont les directions, deux à deux orthogonales, des axes principaux d'inertie au point O. Si la plaque est isocèle ($\tau = 1$), la bissectrice issue de O est axe de symétrie : c'est bien un axe principal d'inertie.

3.3 Rotations

Dans l'espace euclidien E, donnons-nous un repère orthonormé $(O;\vec{i},\vec{j},\vec{k})$ d'origine O et une droite D passant par O, dirigée par le vecteur unitaire $\vec{n} = a\,\vec{i} + b\,\vec{j} + c\vec{k}$ (donc $a^2+b^2+c^2 = 1$).



Pour tout point $M \in E$, notons H le projeté de M sur D et K le projeté de M sur le plan orthogonal à D et passant par O. Le vecteur $\vec{v} = \vec{n} \wedge \overrightarrow{OK}$ est orthogonal à $\vec{n}$ et à $\overrightarrow{OK}$. Comme on a $\|\,\vec{v}\,\| = \|\,\vec{n}\,\|\,\|\,\overrightarrow{OK}\,\| = \|\,\overrightarrow{OK}\,\|$, la rotation d'axe D et d'angle $\pi/2$ transforme $\overrightarrow{OK}$ en $\vec{v}$.

Il s'ensuit que la rotation r d'axe D et d'angle θ transforme $\overrightarrow{OK}$ en $(\cos\theta)\overrightarrow{OK} + (\sin\theta)\vec{v}$.

En posant $M' = r(M)$, on a donc

$$(*) \qquad \overrightarrow{OM'} = \overrightarrow{OH} + (\cos\theta)\,\overrightarrow{OK} + (\sin\theta)\,\vec{v}\ .$$

Notons (x,y,z) les coordonnées de M dans le repère $(O;\vec{i},\vec{j},\vec{k})$ et posons $X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$.

► Les coordonnées de $\overrightarrow{OH} = (\overrightarrow{OM}\cdot\vec{n})\vec{n}$ sont

$$(ax+by+cz)\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a^2 & ab & ac \\ ba & b^2 & bc \\ ca & cb & c^2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = P_D X\ .$$

► Les coordonnées de $\overrightarrow{OK} = \overrightarrow{OM} - \overrightarrow{OH}$ sont

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} - P_D\begin{bmatrix} x \\ y \\ z \end{bmatrix} = (I_3 - P_D)X\ .$$

► Puisque $\overrightarrow{OH}$ est colinéaire à $\vec{n}$, on a $\vec{n}\wedge\overrightarrow{OH} = \mathbf{0}$, donc $\vec{v} = \vec{n}\wedge\overrightarrow{OK} = \vec{n}\wedge\overrightarrow{OM}$ et les coordonnées de $\vec{v}$ sont

$$\begin{bmatrix} bz-cy \\ cx-az \\ ay-bx \end{bmatrix} = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = VX\ .$$

D'après la relation $(*)$, les coordonnées X' du point $M' = r(M)$ sont données par

$$X' = \left[P_D + (\cos\theta)(I_3 - P_D) + (\sin\theta)\, V\right] X\,,$$

autrement dit la matrice de la rotation r dans le repère $(O; \vec{i}, \vec{j}, \vec{k})$ est

$$R = \begin{bmatrix} a^2 & ab & ac \\ ba & b^2 & bc \\ ca & cb & c^2 \end{bmatrix} + (\cos\theta) \begin{bmatrix} 1-a^2 & -ab & -ac \\ -ba & 1-b^2 & -bc \\ -ca & -cb & 1-c^2 \end{bmatrix} + (\sin\theta) \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}.$$

Puisqu'on a $\| \overrightarrow{OM'} \| = \| \overrightarrow{OM} \|$, la matrice R est orthogonale.
Si la rotation r est d'axe $\vec{k}$, alors $(a,b,c) = (0,0,1)$ et la matrice est simplement $A(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$. On voit ainsi que le déterminant de r est égal à 1, donc toute matrice de rotation est de déterminant 1.

Axe et angle d'une rotation. Soit r une rotation de l'espace et soit R sa matrice dans un repère orthonormé dont l'origine est sur l'axe de rotation.

- La rotation laisse fixe les vecteurs de l'axe : tout vecteur de l'axe est donc un vecteur propre de r pour la valeur propre 1.
- Puisque les matrices R et $A(\theta)$ représentent toutes eux la rotation r, ces matrices ont la même trace, donc $1 + 2\cos\theta = \operatorname{tr} R$. Connaissant une matrice de rotation R, on en déduit ainsi immédiatement le cosinus de l'angle.

4. Application à l'analyse de données

On dispose de données numériques sur des entités appelées *individus* : un individu peut être une catégorie socio-professionnelle, une ville, un magasin, etc. Les données portent sur une même liste de *caractères* communs à tous les individus.

Exemple. Voici un tableau des prix pratiqués en euro pour les trois mêmes articles A, B, C dans trois régions différentes R_1, R_2, R_3.
Les individus sont ici les régions. La première colonne du tableau est réservée au premier caractère : elle est formée des prix pratiqués pour l'article A dans les différentes régions. Chaque colonne du tableau est ainsi relative à un caractère : une colonne est un vecteur-caractère.

	A	B	C
R_1	5,7	9,5	4,1
R_2	6,1	8,2	3,5
R_3	7,0	7,9	2,9

Supposons qu'il y a p caractères et n individus avec $n > p$. Comme dans l'exemple, formons la matrice des données.

- Chaque colonne correspond à un caractère : la i-ième colonne est formée des valeurs du i-ème caractère chez les différents individus. Les vecteurs-colonnes $C_1, C_2, \ldots, C_p$, ou *vecteurs-caractères*, sont dans l'espace $\mathbb{R}^n$.

► Chaque ligne correspond à un individu : les vecteurs-ligne sont de la forme ${}^tX_1, {}^tX_2, \dots, {}^tX_n$, où $X_1, X_2, \dots, X_n$ sont des vecteurs de l'espace $\mathbb{R}^p$. Ces vecteurs $X_1, \dots, X_n$ sont les *vecteurs-individu*.

$$\begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array} = [\, C_1 \;\; C_2 \;\; \cdots \;\; C_p \,] = \begin{bmatrix} {}^tX_1 \\ {}^tX_2 \\ \vdots \\ {}^tX_n \end{bmatrix}$$

Représentons chaque vecteur-individu par un point dans $\mathbb{R}^p$: on obtient un *nuage d'individus* formé de n points. L'analyse de données se propose d'étudier la forme de ce nuage pour en déduire des corrélations entre caractères et découvrir des groupes d'individus dont les caractères s'opposent ou se ressemblent.

Préparation des données. Afin de réaliser une étude globale, on introduit la valeur moyenne de chaque caractère et l'on s'intéresse aux écarts à ce caractère moyen. De plus, pour pouvoir comparer des données dont la dispersion dépend des unités employées, on opère également une normalisation sur les écarts à la moyenne. Ainsi, on commence par faire subir aux données deux opérations de régularisation.

Centrage

Pour chaque vecteur-caractère $C_k \in \mathbb{R}^n$, notons m_k la moyenne de ses coordonnées : $m_k = \frac{1}{n}(x_{1k} + x_{2k} + \cdots + x_{nk})$, et définissons le caractère centré $\overline{C}_k$ en posant :

$$\overline{C}_k = \begin{bmatrix} x_{1k} - m_k \\ x_{2k} - m_k \\ \vdots \\ x_{nk} - m_k \end{bmatrix}$$

Puisque $\sum_{i=1}^n (x_{ik} - m_k) = (\sum_{i=1}^n x_{ik}) - n\, m_k = 0$, la somme des coordonnées de $\overline{C}_k$ est nulle.

Normalisation

On définit la variance $\operatorname{var}(C_k)$ et l'écart-type σ_k (voir page 70) :

$$\operatorname{var}(C_k) = \frac{1}{n}\|\overline{C}_k\|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - m_k)^2 \quad \text{et} \quad \sigma_k = \sqrt{\operatorname{var}(C_k)}\ , \quad \text{pour } k = 1, 2, \dots, p.$$

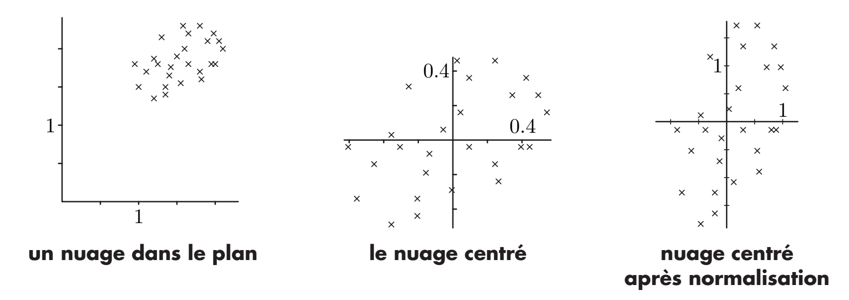
Le caractère centré et normalisé est $D_k = \frac{1}{\sqrt{n}\,\sigma_k}\overline{C}_k = \frac{1}{\|\overline{C}_k\|}\overline{C}_k$.

En posant $y_{ik} = \frac{x_{ik} - m_k}{\sqrt{n}\,\sigma_k}$ pour $1 \leqslant i \leqslant n$ et $1 \leqslant k \leqslant p$, la nouvelle matrice des données est

$$M = \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & \cdots & D_p \end{array} \\ \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} \end{array} \begin{array}{c} \\ {}^tY_1 \\ {}^tY_2 \\ \vdots \\ {}^tY_n \end{array}$$

Considérons les vecteurs-individus $Y_1, \ldots, Y_n$. Pour tout k, leurs coordonnées d'indice k constituent la k-ième colonne de M et comme les caractères sont centrés, on a $y_{1k} + y_{2k} + \cdots + y_{nk} = 0$. Ainsi $Y_1 + Y_2 + \cdots + Y_n = 0$, autrement dit :

le nuage des individus $Y_1, \ldots, Y_n$ a pour centre de gravité l'origine.



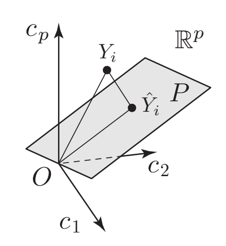
un nuage dans le plan **le nuage centré** **nuage centré après normalisation**

Nous considérons maintenant le nuage des individus $Y_1, \ldots, Y_n$ dans l'espace $\mathbb{R}^p$.

Projection du nuage des individus

Si par exemple chacun des deux premiers caractères est à peu près le même pour tous les individus, tout le nuage d'individus se projette à peu près en un même point dans le plan des deux premières coordonnées de $\mathbb{R}^p$: moins le nuage est dispersé dans un plan de caractères, plus les individus se ressemblent pour les caractères considérés.

On cherche à projeter orthogonalement le nuage sur un plan P de $\mathbb{R}^p$ de manière à perdre le moins d'information possible sur la forme du nuage. Appelons $\hat{Y}_i$ le projeté de Y_i. Selon la méthode des moindres carrés, il s'agit de trouver P pour que la somme $\sum_{i=1}^{n} \|Y_i - \hat{Y}_i\|^2$ soit minimale.



D'après le théorème de Pythagore, on a $\|Y_i\|^2 = \|\hat{Y}_i\|^2 + \|Y_i - \hat{Y}_i\|^2$: il revient donc au même de rendre maximum la somme $s = \sum_{i=1}^{p} \|\hat{Y}_i\|^2$, c'est-à-dire la dispersion des projetés.

La matrice des covariances. Soit U un vecteur-colonne unitaire dans $\mathbb{R}^p$. Pour tout vecteur $Y \in \mathbb{R}^p$, le projeté orthogonal de Y sur la droite engendrée par U est $p(Y) = (Y \cdot U)U$ et l'on a $\|p(Y)\| = |Y \cdot U|$. Puisque $Y \cdot U = ({}^tY)U = ({}^tU)Y$, il vient

$$(*) \qquad s = \sum_{i=1}^{n} |Y_i \cdot U|^2 = \sum_{i=1}^{n} ({}^tU)Y_i({}^tY_i)U = ({}^tU)\Big(\sum_{i=1}^{n} Y_i({}^tY_i)\Big)U$$

Puisque Y_i est un vecteur-colonne de $\mathbb{R}^p$, la matrice $Y_i{}^tY_i$ est carrée de taille p. Par exemple, dans le cas de deux caractères et de trois individus ${}^tY = [\, y_1 \ y_2 \,]$,

${}^tZ = [\,z_1\ \ z_2\,]$ et ${}^tT = [\,t_1\ \ t_2\,]$, on a $M = \begin{bmatrix} y_1 & y_2 \\ z_1 & z_2 \\ t_1 & t_2 \end{bmatrix}$ et

$$Y({}^tY) + Z({}^tZ) + T({}^tT) = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} [\,y_1\ \ y_2\,] + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} [\,z_1\ \ z_2\,] + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} [\,t_1\ \ t_2\,]$$
$$= \begin{bmatrix} y_1^2+z_1^2+t_1^2 & y_1y_2+z_1z_2+t_1t_2 \\ y_2y_1+z_2z_1+t_2t_1 & y_2^2+z_2^2+t_2^2 \end{bmatrix} = ({}^tM)M$$

Dans la matrice $({}^tM)M$, le coefficient en position i-ème ligne, j-ième colonne est $({}^tD_i)D_j$, c'est-à-dire le produit scalaire dans $\mathbb{R}^n$ des vecteurs-caractères D_i et D_j.

Définition

Si M est la matrice des données centrées et normalisées, la matrice $\boldsymbol{C} = ({}^tM)M$ s'appelle la *matrice des covariances*. Elle est carrée de taille p et symétrique.

- Dans la matrice des covariances, tous les coefficients sont compris entre -1 et 1 et tous les coefficients diagonaux sont égaux à 1.
- Les valeurs propres de $\boldsymbol{C}$ sont positives ou nulles et leur somme est égale à p.

En effet, les coefficients de $\boldsymbol{C}$ sont les produits scalaires $({}^tD_i)D_j$. Un produit scalaire est inférieur ou égal au produit des normes et les vecteurs D_i sont de norme 1, donc les coefficients de $\boldsymbol{C}$ sont compris entre -1 et 1 ; pour les coefficients diagonaux, on a $({}^tD_i)D_i = \|D_i\|^2 = 1$. D'après la proposition page 217, les valeurs propres de $\boldsymbol{C}$ sont des réels positifs ou nuls. Leur somme est la trace de $\boldsymbol{C}$ qui, d'après ce qui précède, vaut p (page 174).

Le produit scalaire de deux vecteurs de norme 1 s'interprète comme une corrélation entre ces vecteurs : c'est pourquoi l'on dit aussi que $\boldsymbol{C}$ est la *matrice des corrélations*.

Composantes principales

D'après $(*)$, on a $s = ({}^tU)\,\boldsymbol{C}\,U$ et puisque $\boldsymbol{C}$ est une matrice symétrique, la somme s est maximum lorsqu'on choisit pour U un vecteur propre associé à la plus grande valeur propre de $\boldsymbol{C}$ (proposition page 218 et la remarque qui la suit).
Une telle direction U s'appelle une *composante principale* pour les données : si l'on projette le nuage d'individus sur la droite engendrée par U, la dispersion des points projetés reflète au mieux celle du nuage d'individus.
Considérons les deux plus grandes valeurs propres λ et μ de $\boldsymbol{C}$ et supposons $\lambda > \mu$. Soient U et V des vecteurs propres associés. Puisque la matrice des covariances est symétrique, U et V sont orthogonaux dans $\mathbb{R}^p$. En projetant le nuage d'individus sur le plan P engendré par U et V, on obtient un nuage plan qui ressemble au nuage initial du point de vue des positions relatives des individus.
Cette transformation du nuage s'appelle *l'analyse en composantes principales*. Le plan P s'appelle *le premier plan factoriel*. Il reste à calculer les projections des vecteurs-individus sur ce plan P.
Choisissons les vecteurs propres U et V de norme 1, de sorte que (U, V) est une

base orthonormée de P. Le projeté sur P d'un vecteur $Y \in \mathbb{R}^p$ est $\hat{Y} = aU + bV$, où $a = Y \cdot U = ({}^tY)U$ et $b = Y \cdot V = ({}^tY)V$. Formons la matrice $[U \ \ V]$ qui a p lignes et deux colonnes. Il vient

$$M\,[U \ \ V] = \begin{bmatrix} {}^tY_1 \\ {}^tY_2 \\ \vdots \\ {}^tY_n \end{bmatrix} [U \ \ V] = \begin{bmatrix} ({}^tY_1)U & ({}^tY_1)V \\ ({}^tY_2)U & ({}^tY_2)V \\ \vdots & \vdots \\ ({}^tY_n)U & ({}^tY_n)V \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{bmatrix}$$

Dans la base (U, V), l'individu-projeté $\hat{Y}_i$ a pour coordonnées (a_i, b_i).

La i-ième ligne de la matrice $M\,[U \ \ V]$ est formée des coordonnées du projeté du i-ième individu dans la base orthonormée (U, V) du premier plan factoriel.

Interprétation de la représentation

Dans le premier plan factoriel, les projetés $\hat{Y}_1, \hat{Y}_2, \dots \hat{Y}_n$ des individus forment un nuage : l'analyse de données consiste à découvrir des ressemblances ou des oppositions entre individus en observant la disposition de leurs projetés dans le plan factoriel.

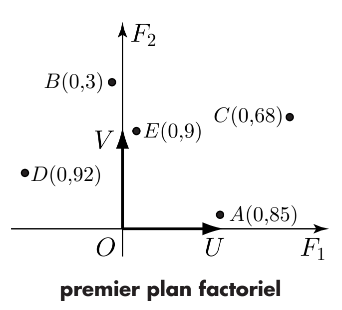
Qualité de la représentation

- Pour que la représentation plane soit exploitable, les valeurs propres λ et μ relatives aux composantes principales doivent être assez grandes par rapport à la somme p des valeurs propres : autrement dit, les *poids* λ/p *et* μ/p doivent être suffisamment grands.
- Des individus très éloignés peuvent avoir des projetés voisins. C'est pourquoi l'on introduit, pour chaque individu, un indice de qualité de sa représentation en projection. La *qualité de représentation* d'un individu Y_i est $\|\hat{Y}_i\|^2/\|Y_i\| = (\cos\theta)^2$, où θ est l'écart angulaire entre le vecteur Y_i et son projeté dans le plan de représentation.

  Si un individu a une qualité de représentation proche de 1, il est proche du plan de représentation et sa distance à l'origine est à peu près la même dans le nuage projeté et dans le nuage initial. De même, la distance entre deux points projetés ne traduit bien leur distance réelle que si leur qualité de représentation est assez grande. Puisque la dissemblance entre deux individus Y_i et Y_j se mesure par leur distance euclidienne $\|Y_i - Y_j\|$, la ressemblance ou la dissemblance entre individus de bonne qualité se lit dans le premier plan factoriel.

- Chacun des vecteurs U et V s'interprète comme une combinaisons de caractères : si par exemple les coordonnées de U sont $[u_1, u_2, \dots, u_p]$, ce vecteur représente un « caractère fictif » ayant un taux de corrélation u_1 avec le premier caractère, u_2 avec le deuxième, u_k avec le k-ième. Un des objectifs de l'analyse est d'interpréter U et V au moyen des caractères définis initialement.

La figure ci-contre montre dans le pan U,V les projetés de quelques individus ; le nombre entre parenthèses est la qualité de représentation. Les points A,D,E sont de bonne qualité : leur distance à l'origine et leurs distances mutuelles traduisent fidèlement celles des individus correspondants. L'individu qui se projette en A a un profil voisin de U et celui qui se projette en E a un profil voisin de V. L'individu qui se projette en D est composé de $-U$ à 64% et de V à 36%. La projection C, de qualité moyenne, n'est guère significative en l'absence d'autres renseignements. Il en va de même pour le point B qui, bien que très proche de l'axe V, est de mauvaise qualité.



premier plan factoriel

Projection des caractères : le second plan factoriel

Rappelons que les vecteurs U et V sont des vecteurs propres de norme 1 associés aux deux plus grandes valeurs propres λ et μ de la matrice des covariances $\boldsymbol{C}$, avec $\lambda > \mu$. Chaque ligne de M est un profil des p caractères. Le caractère fictif $U \in \mathbb{R}^p$ est déterminé par les produits scalaires avec ces n profils, c'est-à-dire par le vecteur MU de coordonnées $[a_1, a_2, \cdots, a_n]$. On a

$$\|MU\|^2 = {}^t(MU)(MU) = ({}^tU)({}^tM)MU = ({}^tU)\,\boldsymbol{C}\,U = ({}^tU)\lambda U = \lambda\|U\|^2 = \lambda$$

donc $\|MU\| = \sqrt{\lambda}$ et de même $\|MV\| = \sqrt{\mu}$. Posons

$$F = \frac{1}{\sqrt{\lambda}} MU \quad \text{et} \quad G = \frac{1}{\sqrt{\mu}} MV.$$

Les vecteurs F et G appartiennent à $\mathbb{R}^n$ et sont de norme 1. Montrons qu'ils sont orthogonaux. On a en effet

$$\sqrt{\lambda\mu}({}^tF)G = {}^t(MU)(MV) = ({}^tU)({}^tM)MV = ({}^tU)\,\boldsymbol{C}\,V = ({}^tU)\mu V = \mu({}^tU)V = 0\,,$$

car U et V sont orthogonaux.

Le plan engendré par les vecteurs F et G s'appelle *le second plan factoriel*. Les vecteurs (F,G) forment une base orthonormée du second plan factoriel.

Pour analyser les caractères $D_1, \ldots, D_p$, projetons-les dans le second plan factoriel. Le projeté D'_k de D_k a pour coordonnées dans la base (F,G) les produits scalaires $\big(({}^tD_k)F, ({}^tD_k)G\big)$, c'est-à-dire la k-ième ligne de la matrice $({}^tM)[\,F\ \ G\,]$ qui a p lignes et deux colonnes. On a $({}^tM)F = \frac{1}{\sqrt{\lambda}}({}^tM)MU = \frac{1}{\sqrt{\lambda}}\lambda U = \sqrt{\lambda}U$, d'où le résultat suivant.

Si $U = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix}$ *et* $V = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix}$, *le vecteur-caractère* D_k *se projette au point* D'_k *de coordonnées* $(\sqrt{\lambda}\,u_k, \sqrt{\mu}\,v_k)$ *dans la base* (F,G).

Puisque D_k a pour norme 1, on a $\|D'_k\| \leqslant 1$. Dans le second plan factoriel, $D'_1, \ldots, D'_p$ forment ainsi un nuage de points, tous situés à une distance au plus 1 de l'origine.

Dans le second plan factoriel, les projetés des caractères sont confinés dans le disque de rayon 1 *centré à l'origine.*

La qualité de représentation d'un caractère D_k se mesure par la distance de D'_k à l'origine : en effet, on a $\|D'_k\|^2 + \|D_k - D'_k\|^2 = \|D_k\|^2 = 1$, donc si $\|D'_k\|$ est peu inférieur à 1, alors $\|D_k - D'_k\|$ est petit et D'_k est proche de D_k.



second plan factoriel

Des projections groupées et situées près du bord du disque traduisent des caractères bien corrélés : c'est pourquoi le cercle unité s'appelle le *cercle des corrélations*. En repérant des groupes de points proches du cercle des corrélations, on découvre des oppositions et des corrélations positives entre les caractères initiaux.

L'analyse en composantes principales utilise conjointement les représentations dans chacun des plans factoriels. Il importe de ne pas confondre ces deux représentations qui ne se situent pas dans le même espace : il s'agit de $\mathbb{R}^p$ pour le premier, de $\mathbb{R}^n$ pour le second. Mais on doit croiser les informations tirées de chaque représentation.

Exemple. On considère des données sociologiques et électorales recueillies dans dix-huit villes françaises à l'occasion des élections municipales de mars 2001 : le tableau des données figure en annexe, page 578. Il y a huit caractères :

poptop est la population totale,

popetr est la population étrangère,

logsoc est le parc de logement social,

chomage est le taux de chomage (d'après l'INSEE),

txhab est le montant de la taxe d'habitation,

revenu est le revenu annuel moyen par habitant (sur la base des revenus en 1998),

votants est le nombre de votants au premier tour,

maj est le pourcentage obtenu par la liste majoritaire.

Les autres tableaux donnent la moyenne des caractères, leur écart-type, la matrice des covariances, les valeurs propres, leur poids, les coordonnées des projections des villes dans le premier plan factoriel et leur qualité de représentation.

Le deux plus grandes valeurs propres ont pour poids respectif 45% et 30% environ : le premier plan factoriel a ainsi un poids d'environ 75%, ce qui justifie une analyse sur les deux premières composantes principales. La figure 1 ci-dessous montre les projections des villes dans le premier plan factoriel.

Dans le second plan factoriel (figure 2), les caractères bien représentés sont ceux qui sont assez proches du cercle des corrélations : c'est le cas du revenu, de la population totale, de la population étrangère, du parc de logement social, du nombre de votants

et du pourcentage obtenu par la liste majoritaire.

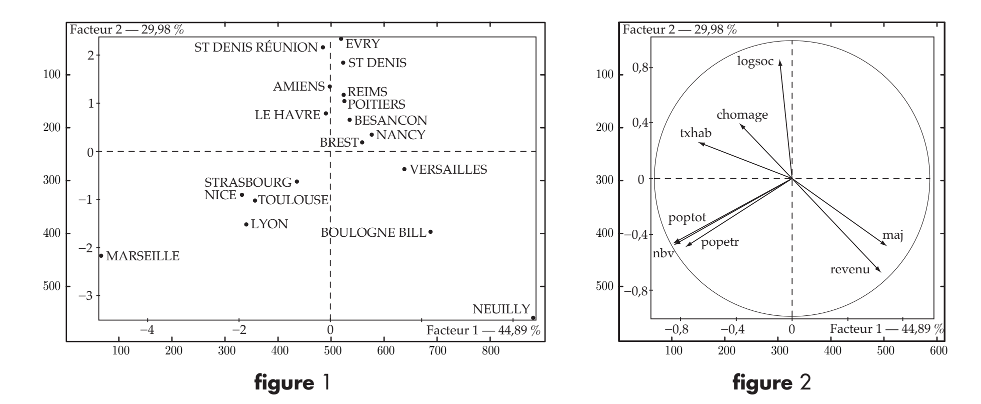


figure 1 **figure** 2

Analyse succincte

a) Dans le second plan factoriel, l'axe F se qualifie par une opposition entre le revenu d'une part et le groupe de caractères population totale, population étrangère et nombre de votants, d'autre part. Le caractère logement social est le seul à qualifier correctement l'axe G : il s'oppose au revenu qui est assez bien corrélé à ce second axe.

b) Dans le premier plan factoriel, la seule ville bien représentée sur l'axe U est Marseille. Amiens et Evry sont bien représentées sur l'axe V.
Marseille et Toulouse se caractérisent par une forte population totale et étrangère, par contraste avec Boulogne et Neuilly qui sont des villes à fort revenus (ces quatre villes ont de bonnes qualités de représentation, supérieures à 0,9).
Evry et Amiens se caractérisent par un important parc de logement social : elles contribuent fortement à la qualification de l'axe V.

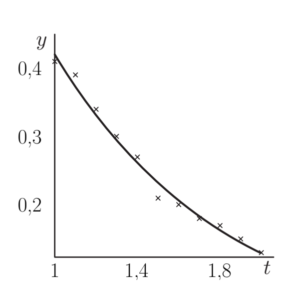
Pour une analyse plus signifiante, il faut davantage de données.

Exercices

1. **Une application de la méthode des moindres carrés.** Une substance diffuse à travers une membrane selon la loi d'action de masse : la quantité non diffusée au temps t s'écrit $y(t) = y_0 e^{-kt}$, où y_0 est la quantité initiale et k un coefficient positif. Le temps t est compté en heures. On veut estimer la quantité y_0 et le coefficient k. On laisse s'établir la diffusion pendant une heure et, pendant l'heure suivante, on mesure toutes les six minutes la quantité de substance non diffusée ; ces mesures s'effectuent donc aux instants $t_0 = 1$, $t_1 = 1{+}0{,}1$, $t_2 = 1{+}0{,}2, \dots, t_9 = 1{+}0{,}9$, $t_{10} = 2$.

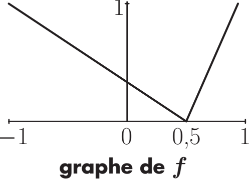
Voici les résultats, y_i désignant la mesure de $y(t_i)$:

t_i	1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2
y_i	0,41	0,39	0,34	0,30	0,27	0,21	0,20	0,18	0,17	0,15	0,13



Puisque $\ln y(t) = -kt + \ln y_0$, on cherche la droite de régression pour les points $(t_i, \ln y_i)$. Montrer que l'on trouve $k = 1{,}17$ et $y_0 = 1{,}36$. Au bout de combien de minutes la moitié de la substance avait-elle diffusé ? La figure ci-contre montre les points mesurés et la courbe $y(t)$ sur l'intervalle $[1, 2]$.

2. Exemple d'approximation polynomiale. On veut approcher par des polynômes de degré au plus 2 la fonction f dont le graphe est représenté ci-contre.



graphe de f

Précisément, on cherche un polynôme $P(x) = a + bx + cx^2$ qui rend minimum l'intégrale $I(Q) = \int_{-1}^{1} \left[f(t) - Q(t)\right]^2 dt$ quand Q parcourt les polynômes de degré au plus 2. Pour cela, considérons l'espace vectoriel des fonctions continues sur $[-1, 1]$ muni du produit scalaire $u \cdot v = \int_{-1}^{1} u(t)v(t)\, dt$ et le sous-espace W engendré par les polynômes $1, x, x^2$.

a) Calculer $f(x)$ pour $-1 \leqslant x \leqslant 1/2$ et pour $1/2 \leqslant x \leqslant 1$.

b) Au moyen de la méthode de Gram-Schmidt appliqué aux polynômes $1, x, x^2$, trouver des polynômes q_0, q_1, q_2 de degrés respectifs $0, 1, 2$ formant une base orthonormée de W.

c) Montrer que P est le projeté orthogonal de f sur W et calculer les coordonnées de P dans la base (q_0, q_1, q_2). Que vaut la distance $d(f, W) = \left[\int_{-1}^{1} \left[f(t) - P(t)\right]^2 dt\right]^{1/2}$ de f au sous-espace W ? Sur un même dessin, représenter le graphe de f et celui de P.

On trouve $P = (1/64)(17 - 16t + 45t^2)$ et $d(f, W) = \sqrt{38}/32 < 2/10$.

@ **3.** Dans l'exemple page 219, l'ellipsoïde $\mathscr{E}_K$ a pour équation $3x^2 + 3y^2 + 2z^2 + 2xz + 2yz = K^2$, ou encore $u^2 + 3v^2 + 4w^2 = K^2$ en utilisant les coordonnées u, v, w dans les axes orthonormés de $\mathscr{E}_K$. Si r est un nombre positif, on note B_r la boule de centre l'origine et de rayon r : un point de coordonnées (x, y, z) est dans B_r si et seulement si $x^2 + y^2 + z^2 \leqslant r^2$.

a) Montrer que dans les coordonnées u, v, w, l'équation de B_r est $u^2 + v^2 + w^2 \leqslant r^2$.

b) Montrer que pour tous nombres u, v, w, on a $u^2 + v^2 + w^2 \leqslant u^2 + 3v^2 + 4w^2 \leqslant 4(u^2 + v^2 + w^2)$ et que les cas d'égalité sont possibles.

c) En déduire que $B_{K/2}$ est la plus grande boule centrée à l'origine et contenue à l'intérieur de $\mathscr{E}_K$ et que B_K la plus petite boule contenant $\mathscr{E}_K$.

@ **4. Dessiner une ellipse.** On pose $f(x, y) = 6x^2 + 4xy + 9y^2$.

a) Écrire la matrice symétrique S telle que $f(x, y) = ({}^tX)SX$ pour tout $X = \begin{bmatrix} x \\ y \end{bmatrix}$.

b) Trouver les axes de l'ellipse $\mathscr{E}$ d'équation $6x^2 + 4xy + 9y^2 = 40$: ils sont dirigés par les vecteurs propres de la matrice S.

c) Calculer des vecteurs E_1 et E_2 formant une base orthonormée de vecteurs propres de S (le vecteur E_1 étant relatif à la plus petite des deux valeurs propres). Montrer que dans le repère d'origine $O = (0, 0)$ et d'axes E_1, E_2, l'ellipse $\mathscr{E}$ a pour équation $5u^2 + 10v^2 = 40$. Dessiner cette ellipse en faisant figurer les axes Ox, Oy et les vecteurs E_1, E_2.

d) En s'inspirant de l'exercice précédent, montrer que

i) le maximum de $x^2 + y^2$ lorsque $6x^2 + 4xy + 9y^2 \leqslant 40$ est égal à 8,

ii) le maximum de $6x^2 + 4xy + 9y^2$ lorsque $x^2 + y^2 \leqslant k^2$ est égal à $10k^2$.

@ **5. Produits scalaires généraux.** Soit S une matrice à coefficients réels, symétrique et définie positive.

a) Montrer qu'il existe une matrice diagonale Δ à coefficients diagonaux strictement positifs et une matrice orthogonale P telle que $S = P^{-1}\Delta^2 P$.

b) On pose $M = \Delta P$. Montrer que la matrice M est inversible et que $S = ({}^tM)M$.

c) Montrer que le produit scalaire $({}^tX)SX'$ est le produit scalaire euclidien usuel des vecteurs MX et MX'.

6. Soient a, b, c, d des nombres réels tels que $a^2 + b^2 + c^2 + d^2 = 1$. Montrer que la matrice $\begin{bmatrix} a & -b & c & -d \\ b & a & -d & -c \\ c & -d & -a & b \\ d & c & b & a \end{bmatrix}$ est orthogonale.

7. Une pyramide orthocentrique. Dans l'espace euclidien $\mathbb{R}^3$, on considère les points $A = (1, 0, -\sqrt{2}/4)$, $B = (-1/2, \sqrt{3}/2, -\sqrt{2}/4)$, $C = (-1/2, -\sqrt{3}/2, -\sqrt{2}/4)$ et $D = (0, 0, 3\sqrt{2}/4)$.

a) Montrer que ces points sont les sommets d'une pyramide régulière (c'est-à-dire un tétraèdre ayant toutes ses arêtes de la même longueur). Montrer que cette pyramide est centrée à l'origine.

b) Quel est l'écart angulaire entre deux arêtes passant par un même sommet ?

c) Montrer que des arêtes opposées (comme AB et CD) sont orthogonales.

@ **8. Rotations et symétries.** Dans l'espace euclidien de dimension 3, on choisit un repère orthonormé $(O; \vec{i}, \vec{j}, \vec{k})$.

a) Trouver la matrice de la rotation d'axe $(1, 1, 1)$ et d'angle θ. Que devient cette matrice lorsque $\theta = 2\pi/3$?

b) Quel est le symétrique du point de coordonnées (u, v, w) par rapport au plan d'équation $x - y + 2z = 0$?

c) Quel est le symétrique du point de coordonnées (u, v, w) par rapport au plan d'équation $x - y + 2z = 1$?

9. Soit M une matrice à coefficients réels, à n lignes et p colonnes. On suppose $n \geqslant p$ et que M est de rang p.

a) Montrer que les matrices $({}^tM)M$ et $M({}^tM)$ sont carrées et symétriques, de taille respective p et n. Montrer que $({}^tM)M$ est inversible (considérer la matrice formée des p premières lignes de M).

b) Soit U un vecteur propre de $({}^tM)M$ pour la valeur propre λ. Montrer que $MU \neq 0$ et que MU est un vecteur propre de $M({}^tM)$ pour la valeur propre λ.

@10. Mouvement apparent pendant une rotation

En un point A d'une voie ferrée en courbe, le centre de courbure est en I ; pour un observateur situé dans le train, au passage en A, les points du paysage ont un mouvement apparent de point fixe I. Supposons que le paysage est dans le plan de la voie ferrée et soit $\vec{k}$ un vecteur de norme 1 orthogonal à ce plan : la vitesse apparente d'un point M est $\omega \vec{k} \wedge \overrightarrow{IM}$, avec $\omega = V/IA$ et V la vitesse linéaire du train en A. Sur la figure, $\vec{v}$ est le vecteur vitesse du train en A. Montrer que les points M qui semblent se diriger vers l'observateur sont situés sur un cercle de diamètre IA (sur une photographie prise vers l'intérieur de la courbe, ces points seront les plus nets).

11. Un exemple de sous-espace dont l'orthogonal est réduit au vecteur nul

Soit V l'espace vectoriel des fonctions $x \mapsto f(x)$ continues par morceaux sur $[0, +\infty[$ et négligeables à l'infini devant $1/x$. Si f et g sont de telles fonctions, le produit $f(x)g(x)$ est négligeable à l'infini devant $1/x^2$, donc l'intégrale généralisée $f \cdot g = \int_0^{+\infty} f(t)g(t)\, dt$ existe (page 325). Cela définit un produit scalaire sur V. Considérons le sous-espace W des fonctions qui valent 0 pour x assez grand : par exemple, pour tout nombre a strictement positif, la fonction C_a qui vaut 1 pour $0 \leqslant x \leqslant a$ et 0 pour $x > a$, est un élément de W. Pour toute fonction $f \in V$, la fonction produit fC_a qui à x associe $f(x)C_a(x)$, appartient à W.

a) Montrer qu'il existe des éléments de V qui n'appartiennent pas à W.

b) Montrer que si $f \in V$, alors $f \cdot (fC_a) = \int_0^a f(t)^2\, dt$ pour tout $a > 0$.

c) En déduire que si f est un élément de V orthogonal à W, alors f est la fonction nulle.

Chapitre 8

Des méthodes numériques

1. Norme et conditionnement d'une matrice

Dans ce chapitre, nous utilisons la norme euclidienne usuelle dans les espaces $\mathbb{R}^n$.

1.1 Norme d'une matrice

On a souvent besoin de savoir dans quelle mesure une application linéaire modifie les normes des vecteurs.

- Prenons par exemple l'application linéaire de $\mathbb{R}^p$ dans $\mathbb{R}$ qui à tout vecteur $X \in \mathbb{R}^p$ associe le produit scalaire $C \cdot X$ par un vecteur-colonne $C \in \mathbb{R}^p$ donné. D'après l'inégalité de Cauchy-Schwarz, on a $|C \cdot X| \leqslant \|C\| \, \|X\|$ et l'égalité a lieu si C et X sont colinéaires. Lorsque X parcourt les vecteurs non nuls de $\mathbb{R}^p$, les rapports $\dfrac{|C \cdot X|}{\|X\|}$ ont donc pour valeur maximum $\|C\|$.
- Soit A une matrice à coefficients réels ayant n lignes et p colonnes. Notons $C_1, C_2, \ldots, C_n$ les matrices-ligne de A. Si X est un vecteur de $\mathbb{R}^p$, le i-ème coefficient du vecteur $AX \in \mathbb{R}^n$ est le produit scalaire $C_i \cdot X$, donc $AX = \begin{bmatrix} C_1 \cdot X \\ C_2 \cdot X \\ \vdots \\ C_n \cdot X \end{bmatrix}$.

  En notant de la même manière la norme euclidienne dans $\mathbb{R}^n$ et dans $\mathbb{R}^p$, il vient $\|AX\|^2 = (C_1 \cdot X)^2 + (C_2 \cdot X)^2 + \cdots + (C_n \cdot X)^2 \leqslant \big(\|C_1\|^2 + \|C_2\|^2 + \cdots + \|C_n\|^2\big) \|X\|^2$.

  Quand X parcourt les vecteurs non nuls de $\mathbb{R}^p$, les rapports $\dfrac{\|AX\|^2}{\|X\|^2}$ restent donc inférieurs ou égaux au nombre $\big(N(A)\big)^2 = \|C_1\|^2 + \|C_2\|^2 + \cdots + \|C_n\|^2 = \sum_{i,j} a_{ij}^2$ égal à la somme des carrés des coefficients de A.

Un rapport $\|AX\|/\|X\|$ reste inchangé quand on remplace X par λX : on obtient donc tous ces nombres en faisant seulement parcourir à X les vecteurs de norme 1. Posons $\varphi(X) = \|AX\|/\|X\|$ pour X appartenant à la sphère $S = \{X \in \mathbb{R}^p \mid \|X\| = 1\}$. La fonction φ est continue et S est un domaine compact (page 362), donc φ atteint son maximum en un point de S.

Définition

Soit A une matrice à p colonnes. La *norme* de la matrice A, notée $\|A\|$, est le plus grand des nombres $\dfrac{\|AX\|}{\|X\|}$, quand X parcourt les vecteurs non nuls de $\mathbb{R}^p$.

L'emploi du mot « norme » pour ce nombre se justifie par les propriétés (b), (c) et (d) ci-dessous.

Propriétés de la norme d'une matrice. *Soit* $A \in \mathscr{M}_{n,p}(\mathbb{R})$.

a) *Pour tout vecteur* $X \in \mathbb{R}^p$, *on a* $\|AX\| \leqslant \|A\|\,\|X\|$.
b) $\|A\| = 0 \iff A = 0$.
c) $\|\lambda A\| = |\lambda|\|A\|$ *pour tout nombre réel* λ.
d) *Si* $A' \in \mathscr{M}_{n,p}(\mathbb{R})$, *alors* $\|A + A'\| \leqslant \|A\| + \|A'\|$.
e) *Si* $B \in \mathscr{M}_{p,q}(\mathbb{R})$, *alors* $\|AB\| \leqslant \|A\|\,\|B\|$.

Démonstration. Par définition de la norme de A, on a $\dfrac{\|AX\|}{\|X\|} \leqslant \|A\|$ pour tout vecteur non nul $X \in \mathbb{R}^p$, d'où l'inégalité (a). On en déduit que si $\|A\| = 0$, alors $AX = 0$ pour tout X, donc la matrice A est nulle ; réciproquement, si $A = 0$, alors $\|AX\| = \|0\| = 0$ pour tout X. Si $\lambda \in \mathbb{R}$, alors $\dfrac{\|(\lambda A)X\|}{\|X\|} = |\lambda| \dfrac{\|AX\|}{\|X\|}$, d'où (c). D'après l'inégalité triangulaire (page 204), on a $\|(A + A')X\| = \|AX + A'X\| \leqslant \|AX\| + \|A'X\|$. On en déduit $\|(A + A')X\| \leqslant (\|A\| + \|A'\|)\|X\|$; pour tout vecteur $X \neq 0$, le nombre $\dfrac{\|(A + A')X\|}{\|X\|}$ est donc inférieur ou égal à $\|A\| + \|A'\|$, d'où (d). Si B est une matrice réelle à p lignes et q colonnes, alors pour tout $X \in \mathbb{R}^q$, on a $\|BX\| \leqslant \|B\|\,\|X\|$ et $\|A(BX)\| \leqslant \|A\|\,\|BX\|$, donc $\|(AB)X\| = \|A(BX)\| \leqslant \|A\|\,\|B\|\,\|X\|$, ce qui démontre (e). ■

Calculer la norme d'une matrice est rarement commode. Le plus souvent, on se contente d'une majoration ou d'une estimation numérique. Voici des exceptions où le calcul de la norme est cependant très simple.

Exemples

1) Si A est une matrice-ligne, $\|A\|$ est simplement la norme de A considérée comme un vecteur (voir l'exemple au début du paragraphe).
2) Supposons que la matrice A est orthogonale de taille n. Dans ce cas, la transformation $X \mapsto AX$ de $\mathbb{R}^n$ est une isométrie et $\|AX\| = \|X\|$ pour tout X, donc $\|A\| = 1$.
3) Prenons une matrice diagonale $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$. Si X est un vecteur-colonne de coordonnées $(x_1, x_2, \ldots, x_n)$, les coordonnées de DX sont $(d_1x_1, d_2x_2, \ldots, d_nx_n)$ et $\|DX\|^2 = d_1^2x_1^2 + d_2^2x_2^2 + \cdots + d_n^2x_n^2$. Posons $m = \max\big(|d_1|, |d_2|, \ldots, |d_n|\big)$. On a

$\|DX\|^2 \leqslant m^2x_1^2 + m^2x_2^2 + \cdots + m^2x_n^2 = m^2\|X\|^2$. Les nombres $\dfrac{\|DX\|}{\|X\|}$, où $X \neq 0$, sont tous inférieurs ou égaux à m, donc aussi le plus grand d'entre eux qui est $\|D\|$. Le nombre m est l'un des $|d_i|$: supposons $m = |d_k|$. Pour le vecteur canonique $\boldsymbol{E_k}$, on a $\|\boldsymbol{E_k}\| = 1$, $D\boldsymbol{E_k} = d_k\boldsymbol{E_k}$, donc $\|D\boldsymbol{E_k}\| = |d_k|$. Ainsi $|d_k|$ est le plus grand des quotients $\|DX\|/\|X\|$, autrement dit

$$\|\operatorname{diag}(d_1, d_2, \ldots, d_n)\| = \max\big(|d_1|, |d_2|, \ldots, |d_n|\big)$$

Pour une matrice générale, on a les résultats suivants.

Proposition. *Soit A une matrice à coefficients réels.*

- *La norme de A est la racine carrée de la plus grande valeur propre de la matrice $({}^tA)A$.*
- *Si A possède p colonnes, on a l'encadrement $N(A)/\sqrt{p} \leqslant \|A\| \leqslant N(A)$, où $\big(N(A)\big)^2$ est la somme des carrés des coefficients de A.*

Démonstration. Supposons que A possède p colonnes. Pour tout $X \in \mathbb{R}^p$, on a $\|AX\|^2 = {}^t(AX)(AX) = ({}^tX)({}^tA)AX$. La matrice $S = ({}^tA)A$ est carrée de taille p, symétrique et ses valeurs propres sont des nombres réels positifs ou nuls (proposition page 217). Soit R la plus grande valeur propre de S. D'après la proposition page 218 et la remarque qui la suit, on a $\|AX\|^2 = ({}^tX)SX \leqslant R\|X\|^2$ et l'égalité est obtenue quand X est un vecteur propre de S pour la valeur propre R. Quand X parcourt les vecteurs non nuls de $\mathbb{R}^p$, les rapports $\|AX\|/\|X\|$ ont donc pour maximum $\sqrt{R}$. Nous avons déjà démontré en introduction l'inégalité $\|A\| \leqslant N(A)$. La i-ème colonne A_i de A est $A\boldsymbol{E_i}$, où $\boldsymbol{E_i}$ est le i-ème vecteur canonique, donc $\|A_i\| = \|A\boldsymbol{E_i}\| \leqslant \|A\|\|\boldsymbol{E_i}\| = \|A\|$, car $\boldsymbol{E_i}$ est de norme 1. En élevant au carré et en ajoutant ces inégalités pour $i = 1, 2, \ldots, p$, on obtient $p\|A\|^2 \geqslant \|A_1\|^2 + \|A_2\|^2 + \cdots + \|A_p\|^2 = \big(N(A)\big)^2$, d'où l'inégalité $\|A\| \geqslant N(A)/\sqrt{p}$. ■

Calcul numérique de la norme d'une matrice

Première méthode. Pour avoir une estimation de la norme d'une matrice A à n lignes et p colonnes,

- on génère des vecteurs $X \in \mathbb{R}^p$ au hasard,
- on calcule $Y = AX$ et le rapport $r = \dfrac{\|Y\|}{\|X\|}$,
- et l'on prend le plus grand de ces nombres r comme valeur approchée de $\|A\|$.

Si l'on a pris suffisamment de vecteurs X et s'ils sont assez dispersés, r est une bonne approximation par défaut de la norme de A.

Seconde méthode. Si p n'est pas trop grand, on peut aussi calculer le polynôme caractéristique $f(x)$ de la matrice $({}^tA)A$ et chercher une valeur approchée de la plus grande racine $\lambda_{\max}$ de f ; on a alors une valeur approchée de $\|A\| = \sqrt{\lambda_{\max}}$.
Puisque le polynôme f a toutes ses racines réelles, la méthode de Newton est tout à fait adaptée au calcul de la plus grande racine (voir pages 307 et 309).

Rappelons que la suite de Newton est définie par son premier terme x_0 et la relation $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$. Si l'on choisit x_0 plus grand que $\lambda_{\max}$, par exemple $x_0 = \big(N(A)\big)^2$, alors x_n tend rapidement vers $\lambda_{\max}$ par valeurs supérieures.
La méthode de Hörner (page 48) permet un calcul efficace des nombres $f(x_n)$ et $f'(x_n)$.

Exemple. Soit $A = \begin{bmatrix} -14{,}2 & 13{,}8 & 17{,}4 & 10{,}0 \\ -8{,}6 & 18{,}8 & 17{,}0 & 16{,}0 \\ -19{,}0 & 12{,}2 & -19{,}4 & 12{,}2 \\ 9{,}8 & 10{,}0 & 12{,}4 & 12{,}6 \\ 4{,}4 & 0{,}4 & 10{,}8 & -4{,}8 \end{bmatrix}$.

a) Posons $x_0 = 3311 > \big(N(A)\big)^2$. Dans la suite de Newton pour le polynôme caractéristique de $({}^tA)A$, le terme $x_8 = 2049{,}8363\ldots$ approche la plus grande racine à 10^{-5} près. En prenant la racine carrée, on obtient $\|A\| \simeq 45{,}275118$, valeur approchée par excès à 10^{-7} près.

b) Voici les résultats obtenus avec la première méthode. La première ligne du tableau indique le nombre de vecteurs $X \in \mathbb{R}^4$ générés au hasard et la deuxième ligne donne $r = \max(\|AX\|/\|X\|)$, valeur approchée par défaut de $\|A\|$; dans la troisième ligne, on a calculé l'erreur relative $(\|A\| - r)/\|A\|$.

nombre d'essais	10	25	40	55	70	85	100
valeur approchée de $\|A\|$	39,06	44,14	43,36	43,54	43,95	44,24	44,39
erreur relative	0,13	0,02	0,04	0,04	0,03	0,02	0,02

1.2 Conditionnement d'une matrice

La résolution de certaines équations linéaires $AX = B$ présente des difficultés inattendues : de petites variations sur les coefficients du vecteur B produisent de grandes variations sur la solution X. Cela est très gênant, car dans la pratique numérique, les coefficients de B, et même de A, ne sont en général connus que de manière approchée. Ce type d'erreur s'ajoutant aux inévitables arrondis commis pendant la résolution, la solution calculée n'est plus fiable.

Exemple (d'après R.S. Wilson). Posons

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}, \quad B = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \quad \text{et} \quad dB = \begin{bmatrix} 0{,}1 \\ -0{,}1 \\ 0{,}1 \\ -0{,}1 \end{bmatrix}.$$

► La solution de l'équation linéaire $AX = B$ est $X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$,

► et si l'on pose $A(X + dX) = B + dB$, on obtient $dX = \begin{bmatrix} 9{,}2 \\ -12{,}6 \\ 4{,}5 \\ -1{,}1 \end{bmatrix}$.

La petite variation relative $vB = \dfrac{\|dB\|}{\|B\|}$ qui vaut environ $3{,}3.10^{-3}$ provoque dans la solution une variation relative $vX = \dfrac{\|dX\|}{\|X\|}$ d'environ 8,2, soit un taux vX/vB supérieur à 2450.

Donnons-nous une matrice carrée A inversible de taille n et un vecteur-colonne $B \in \mathbb{R}^n$. Soit X la solution de l'équation linéaire $AX = B$. Modifions B en $B + dB$, où $dB \in \mathbb{R}^n$: la solution devient $X + dX$, telle que $A(X + dX) = B + dB$.
On a $A(dX) = dB$, donc $dX = A^{-1}(dB)$ et $\|dX\| \leqslant \|A^{-1}\|\|dB\|$. En multipliant par l'inégalité $\|B\| \leqslant \|A\|\|X\|$, il vient $\|B\|\|dX\| \leqslant \|A\|\|A^{-1}\|\|X\|\|dB\|$, d'où

$$\frac{\|dX\|}{\|X\|} \leqslant \|A\|\|A^{-1}\|\frac{\|dB\|}{\|B\|}$$

Définition

Soit A une matrice carrée inversible de taille n. Le nombre $\|A\|\|A^{-1}\|$ s'appelle le *conditionnement* de A et se note $\operatorname{cond}(A)$.

- Pour une matrice diagonale inversible $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$, le conditionnement est le quotient $\operatorname{cond}(D) = \dfrac{\max |d_i|}{\min |d_i|}$.
- Puisque $I_n = A(A^{-1})$, il vient $1 = \|I_n\| = \|A(A^{-1})\| \leqslant \|A\|\|A^{-1}\|$ d'après les propriétés de la norme. Le conditionnement est donc toujours supérieur ou égal à 1.
- On a l'égalité $\operatorname{cond}(A) = \operatorname{cond}(A^{-1})$.

Proposition. *Soit A une matrice inversible de taille n et soient B et dB des vecteurs-colonne de $\mathbb{R}^n$. Si $AX = B$ et $A(X + dX) = B + dB$, alors* $\dfrac{\|dX\|}{\|X\|} \leqslant \operatorname{cond}(A)\dfrac{\|dB\|}{\|B\|}$.

Calcul du conditionnement

On peut calculer les normes $\|A\|$ et $\|A^{-1}\|$ comme au précédent paragraphe. Dans l'exemple précédent, on trouve ainsi que le conditionnement vaut environ 2984 (valeur compatible avec les rapports $\dfrac{\|dX\|}{\|X\|}$ et $\dfrac{\|dB\|}{\|B\|}$ obtenus).
On a aussi la formule suivante.

Proposition. *Soient $\lambda_{\max}$ la plus grande valeur propre de la matrice $({}^tA)A$ et $\lambda_{\min}$ la plus petite. Alors* $\operatorname{cond}(A) = \sqrt{\lambda_{\max}/\lambda_{\min}}$.

Démonstration. Posons $S = ({}^tA)A$ et rappelons que toutes les valeurs propres de S sont strictement positives, car A est inversible (proposition page 217). Pour toute matrice inversible P de taille n, la matrice $P^{-1}SP$ a mêmes valeurs propres que S. En choisissant $P = {}^tA$, cela montre que $\lambda_{\min}$ est la plus petite valeur propre de $P^{-1}SP = A\,{}^tA$. Les valeurs propres de l'inverse d'une matrice sont les inverses des valeurs propres de cette matrice, donc $1/\lambda_{\min}$ est la plus grande valeur propre de $(A\,{}^tA)^{-1} = {}^t(A^{-1})(A^{-1})$. Par conséquent, on a $\|A^{-1}\| = \sqrt{1/\lambda_{\min}}$. Puisque $\|A\| = \sqrt{\lambda_{\max}}$, le résultat s'ensuit. ■

Soit f le polynôme caractéristique de $({}^tA)A$. L'équation $f(1/x)=0$ a pour solutions les inverses des racines de f et $1/\lambda_{\min}$ est la plus grande de ces solutions. Posons $g(x)=x^n f(1/x)$, où $n=\deg f$ est la taille de la matrice $({}^tA)A$; alors g est un polynôme de degré n et l'on peut calculer $1/\lambda_{\min}$ en appliquant la méthode de Newton à g.

Estimation pratique. La formule de la proposition est valable pour toutes les perturbations dB du second membre, notamment celles qui provoquent la plus grande variation sur la solution X ; mais dans la pratique, un vecteur dB au hasard ne provoquera que des variations dX nettement inférieures au maximum prévu par le conditionnement. On se contente donc souvent d'estimer numériquement un coefficient $C(A)$ tel que l'inégalité $\frac{\|dX\|}{\|X\|} \leqslant C(A)\frac{\|dB\|}{\|B\|}$ soit valable pour suffisamment de second membres B et suffisamment de perturbations dB prises au hasard :

- on génère un assez grand nombre de vecteurs B et dB au hasard,
- on résout les équations $AX = B$ et $A(X + dX) = B + dB$,
- on prend comme valeur de $C(A)$ le plus grand des nombres $\frac{\|dX\|}{\|X\|}\left[\frac{\|dB\|}{\|B\|}\right]^{-1}$.

Le nombre $C(A)$ est en général bien inférieur au conditionnement, mais, en pratique, il n'est pas déraisonnable de l'utiliser pour des calculs d'erreurs.

2. Résolution d'équations linéaires

Dans ce paragraphe, on considère des équations linéaires $AX = B$, où A est une matrice inversible de taille n à coefficients réels ou complexes.

2.1 Factorisation LU

Lorsque la matrice A est triangulaire, l'équation linéaire $AX = B$ se résout facilement : si par exemple A est triangulaire inférieure, la première équation fournit x_1, première coordonnée de la solution $X = (x_1, x_2, \ldots, x_n)$, puis en reportant dans la deuxième équation, on obtient x_2 et ainsi de suite ; si A est triangulaire supérieure, il faut commencer par la dernière équation et remonter jusqu'à la première.

Prenons une matrice A produit de matrices triangulaires, par exemple $A = LU$, où L est triangulaire inférieure et U triangulaire supérieure. Puisque A est inversible, L et U le sont aussi : en effet, $\det A = (\det L)(\det U)$ est non nul, donc L et U ont des déterminants non nuls. Si C est le vecteur tel que $LC = B$, alors on a

$$AX = B \iff LUX = LC \iff UX = C\,, \quad \text{car } L \text{ est inversible.}$$

La résolution de l'équation $AX = B$ peut donc s'opérer en deux étapes plus simples :

- résolution de l'équation triangulaire $LC = B$,
- puis résolution de l'équation triangulaire $UX = C$.

Nous allons montrer que moyennant une hypothèse souvent satisfaite, on peut calculer efficacement des matrices L et U telle que $A = LU$.

Recherche de la factorisation

On dit qu'une matrice A possède une *factorisation LU* s'il existe une matrice L triangulaire inférieure à coefficients diagonaux tous égaux à 1 et une matrice U triangulaire supérieure, telles que $A = LU$ (l'appellation « LU » vient de l'anglais « lower triangular » et « upper triangular »).

Notation : Si M est une matrice carrée d'ordre n et si k est un entier tel que $1 \leqslant k \leqslant n$, notons $M^{(k)}$ la matrice carrée de taille k formée avec les k premières lignes et les k premières colonnes de M.

Ainsi par exemple, pour $L = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix}$ et $U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$, on a

$$L^{(1)} = [\,1\,] \qquad U^{(1)} = [\,u_{11}\,] \qquad L^{(1)}U^{(1)} = [\,u_{11}\,] = (LU)^{(1)}$$

$$L^{(2)} = \begin{bmatrix} 1 & 0 \\ \ell_{21} & 1 \end{bmatrix} \qquad U^{(2)} = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \qquad L^{(2)}U^{(2)} = \begin{bmatrix} u_{11} & u_{12} \\ \ell_{21}u_{11} & \ell_{21}u_{12} + u_{22} \end{bmatrix} = (LU)^{(2)}$$

et plus généralement, pour des matrices L et U de taille quelconque, on a $L^{(k)}U^{(k)} = (LU)^{(k)}$.

Si $A = LU$, alors L et U sont inversibles (car A est inversible), donc leurs coefficients diagonaux sont tous non nuls. Par suite, les matrices $L^{(k)}$ et $U^{(k)}$ sont inversibles et le déterminant de $A^{(k)} = (LU)^{(k)} = L^{(k)}U^{(k)}$ est non nul. Pour que la matrice A se factorise en LU, il faut donc que les déterminants des matrices $A^{(k)}$ soient tous non nuls. La proposition suivante affirme que cette condition est aussi suffisante.

Proposition. *Soit A une matrice inversible. Si toutes les matrices $A^{(k)}$ ont un déterminant non nul, alors A possède une factorisation LU unique.*

Démonstration. Décomposons A sous la forme $A = \begin{bmatrix} A' & P \\ Q & a \end{bmatrix}$, où A' est carrée de taille $n-1$, P est une matrice-colonne à $n-1$ lignes, Q est une matrice-ligne à $n-1$ colonnes et $a \in \mathbb{R}$. Cherchons aussi L et U sous la forme $L = \begin{bmatrix} L' & 0 \\ X & 1 \end{bmatrix}$ et $U = \begin{bmatrix} U' & Y \\ 0 & u \end{bmatrix}$. L'égalité $LU = A$ équivaut à $L'U' = A'$, $L'Y = P$, $XU' = Q$ et $XY + u = a$. Pour montrer l'existence de L et U, raisonnons par récurrence sur la taille de la matrice A. Les matrices $A'^{(1)} = A^{(1)}, A'^{(2)} = A^{(2)}, \ldots, A'^{(n-1)} = A^{(n-1)}$ ayant leur déterminant non nul, la décomposition $A' = L'U'$ existe par hypothèse de récurrence. Les équations $L'Y = P$ et $XU' = Q$ ont une solution car L' et U' sont inversibles et l'on pose $u = a - XY$: cela définit les matrices L et U. Si $A = \tilde{L}\tilde{U}$ est une (autre) factorisation LU, alors la matrice $L^{-1}\tilde{L} = U\tilde{U}^{-1}$ est triangulaire supérieure et triangulaire inférieure avec des 1 sur la diagonale, donc $L^{-1}\tilde{L} = U\tilde{U}^{-1} = I_n$ et $\tilde{L} = L$, $\tilde{U} = U$. ■

Pratique des calculs

On calcule les matrices $L^{(k)}$ et $U^{(k)}$ de proche en proche, jusqu'à $L=L^{(n)}$ et $U=U^{(n)}$:

- $L^{(1)} = [\,1\,]$ et $U^{(1)} = A^{(1)}$;
- Si l'on a calculé $L^{(k)}$ et $U^{(k)}$, on pose comme dans la démonstration précédente :

$$L^{(k+1)} = \begin{bmatrix} L^{(k)} & 0 \\ X & 1 \end{bmatrix}, \quad U^{(k+1)} = \begin{bmatrix} U^{(k)} & Y \\ 0 & u \end{bmatrix} \quad \text{et} \quad A^{(k+1)} = \begin{bmatrix} A^{(k)} & P \\ Q & a \end{bmatrix},$$

  avec a et u des scalaires ; on calcule alors la matrice-colonne X et la matrice-ligne Y en résolvant les équations $L^{(k)}Y = P$, $XU^{(k)} = Q$, puis on obtient le scalaire u au moyen de l'égalité $XY + u = a$.

Pour calculer L et U lorsque A est de grande taille n, le nombre d'opérations à effectuer (additions, multiplications et divisions) est de l'ordre de n^3.

2.2 Méthode de relaxation

Au lieu de résoudre algébriquement l'équation linéaire $AX = B$, on peut chercher à approcher la solution par des vecteurs $X^{(0)}, X^{(1)}, \dots, X^{(k)}, \dots$ qui tendent vers la solution. Voici une méthode couramment pratiquée pour construire une telle suite.

Hypothèse générale : la matrice $A=[a_{ij}]$ est inversible et ses coefficients diagonaux a_{ii} sont tous non nuls.

On note $b_1, b_2, \dots, b_n$ les coordonnées du vecteur B.

Construction des vecteurs $X^{(k)}$

Pour définir une méthode de relaxation, on se donne un nombre ω tel que $0<\omega<2$.

- On choisit un vecteur initial $X^{(0)}$.
- On passe de $X^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})$ à $X^{(k+1)} = (x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)})$ en appliquant la formule suivante pour $i = 1, 2, \dots, n$:

$$(1) \qquad x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)}\Big)$$

  où l'on convient que si $i = 1$, le signe $\sum_{j=1}^{i-1}$ donne comme résultat 0.

Par cette formule, on calcule successivement $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)}$, connaissant $X^{(k)}$. Supposons que les vecteurs $X^{(k)}$ tendent vers une limite X quand k tend vers l'infini. Pour chaque valeur fixée de i, on peut passer à la limite dans l'égalité (1) : en notant $x_1, x_2, \dots, x_n$ les coordonnées de la limite X, on obtient $x_i = x_i + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j=1}^{n} a_{ij}x_j\Big)$, donc $0 = b_i - \sum_{j=1}^{n} a_{ij}x_j$; puisque la somme est la i-ième coordonnée de AX, on a $B = AX$. Cela montre que

si les vecteurs $X^{(k)}$ ont une limite, cette limite est la solution de l'équation $AX = B$.

Algorithme de passage de $X^{(k)}$ à $X^{(k+1)}$

La programmation est très simple :

initialisation : $[x_1, x_2, \ldots, x_n] \leftarrow X^{(k)}$

boucle : pour i de 1 à n, faire

$$\left[\begin{array}{l} s \leftarrow 0 \\ \text{pour } j \text{ de } 1 \text{ à } n, \text{ faire } s \leftarrow s + a_{ij}x_j \\ r \leftarrow b_i - s \\ x_i \leftarrow x_i + \omega r / a_{ii} \end{array}\right.$$

fin : $X^{(k+1)} \leftarrow [x_1, x_2, \ldots, x_n]$

Test d'arrêt

Pour tout k, on calcule le « résidu » $r^{(k)} = B - AX^{(k)}$ et l'on s'arrête si

$$\frac{\|r^{(k)}\|}{\|B\|} < \varepsilon \quad , \quad \text{où } \varepsilon \text{ est une précision choisie.}$$

Notons $e^{(k)} = X - X^{(k)}$ l'erreur commise en s'arrêtant à la k-ième itération.
On a $AX = B$ et $AX^{(k)} = B - r^{(k)}$, donc $Ae^{(k)} = AX - AX^{(k)} = r^{(k)}$ et $\|e^{(k)}\| = \|A^{-1}r^{(k)}\| \leqslant \|A^{-1}\|\|r^{(k)}\|$.
Après un test d'arrêt positif, on a $\|e^{(k)}\| \leqslant \varepsilon \|A^{-1}\|\|B\| \leqslant \varepsilon \|A^{-1}\|\|A\|\|X\|$. Pour l'erreur relative, il vient donc la majoration

$$\frac{\|e^{(k)}\|}{\|X\|} \leqslant \varepsilon \operatorname{cond}(A)$$

La matrice d'itération

Nous allons voir que les vecteurs $X^{(k)}$ sont les itérés de $X^{(0)}$ par une transformation affine de la forme $X \mapsto \mathscr{L}_\omega X + K$, où $\mathscr{L}_\omega$ est une certaine matrice ne dépendant que de A et de ω. Décomposons la matrice A en trois matrices D, E, F :

- la matrice D est la matrice diagonale $\operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ formée des coefficients diagonaux de A (supposés tous non nuls) ;
- la matrice $(-E)$ est la partie triangulaire strictement en dessous de la diagonale ;
- la matrice $(-F)$ est la partie triangulaire strictement au dessus de la diagonale.

$$A = \begin{bmatrix} \ddots & & -F \\ & D & \\ -E & & \ddots \end{bmatrix}$$

La formule (1) s'écrit $X^{(k+1)} = X^{(k)} + \omega D^{-1}\big[B + EX^{(k+1)} + (F - D)X^{(k)}\big]$, ou encore

$$(I_n - \omega D^{-1}E)X^{(k+1)} = \big[(1-\omega)I_n + \omega D^{-1}F\big]X^{(k)} + \omega D^{-1}B$$

En posant
$$L=D^{-1}E=\begin{bmatrix} 0 & & & \\ \frac{-a_{21}}{a_{22}} & \ddots & & 0 \\ \frac{-a_{31}}{a_{33}} & \frac{-a_{32}}{a_{33}} & \ddots & \\ \vdots & & & \ddots \\ \frac{-a_{n1}}{a_{nn}} & \frac{-a_{n2}}{a_{nn}} & \cdots & \frac{-a_{n\,n-1}}{a_{nn}} & 0 \end{bmatrix} \text{ et } U=D^{-1}F=\begin{bmatrix} 0 & \frac{-a_{12}}{a_{11}} & \frac{-a_{13}}{a_{11}} & \cdots & \frac{-a_{1n}}{a_{11}} \\ & \ddots & \frac{-a_{23}}{a_{22}} & \cdots & \frac{-a_{2n}}{a_{22}} \\ & & \ddots & & \vdots \\ & 0 & & \ddots & \frac{-a_{n-1\,n}}{a_{n-1\,n-1}} \\ & & & & 0 \end{bmatrix},$$
il vient $(I_n - \omega L)X^{(k+1)} = \big[(1-\omega)I_n + \omega U\big]X^{(k)} + \omega D^{-1}B$. La matrice triangulaire $I_n - \omega L$ est inversible, car ses coefficients diagonaux sont tous égaux à 1. On a donc
$$X^{(k+1)} = (I_n - \omega L)^{-1}\big[(1-\omega)I_n + \omega U\big]X^{(k)} + \omega(I_n - \omega L)^{-1}D^{-1}B$$
et en posant $\mathscr{L}_\omega = (I_n - \omega L)^{-1}\big[(1-\omega)I_n + \omega U\big]$ et $K = \omega(I_n - \omega L)^{-1}D^{-1}B$, on obtient

(2) $$X^{(k+1)} = \mathscr{L}_\omega X^{(k)} + K\,, \quad \text{pour tout entier } k \geqslant 0.$$

La solution de l'équation linéaire $AX = B$ est le point fixe de cette transformation. D'après le corollaire page 185, on en déduit :

> *pour que la suite de $X^{(k)}$ converge quel que soit le vecteur initial X_0, il faut et il suffit que la matrice d'itération $\mathscr{L}_\omega$ ait toutes ses valeurs propres (réelles ou complexes) de module strictement inférieur à 1.*

On a $(I_n - \omega L)\mathscr{L}_\omega = (1-\omega)I_n + \omega U$ et la matrice triangulaire $I_n - \omega L$ a pour déterminant 1, donc $\det \mathscr{L}_\omega = \det\big[(1-\omega)I_n + \omega U\big] = (1-\omega)^n$. On sait que le déterminant d'une matrice est le produit de toutes les valeurs propres réelles ou complexes. Si les valeurs propres de $\mathscr{L}_\omega$ sont de module inférieur à 1, alors en faisant leur produit, il vient $|1-\omega|^n < 1$, donc $|1-\omega| < 1$; puisque nous avons pris ω réel, cela implique que ω est strictement compris entre 0 et 2. *La condition $0 < \omega < 2$ est nécessaire pour que la méthode de relaxation converge quel que soit le vecteur initial X_0.*

Des conditions suffisantes de convergence

Dans les applications, la matrice A est souvent bien particulière, par exemple symétrique ou tridiagonale (voir page 142) ; il est également fréquent que les coefficients diagonaux soient prépondérants au sens de la définition suivante.

Définition

Une matrice carrée $A = [a_{ij}]$ est *à diagonale strictement dominante* si le module de chaque coefficient diagonal est strictement supérieur à la somme des modules des autres coefficients situés sur la même ligne : pour tout i, $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$.

Proposition. *Toute matrice carrée à diagonale strictement dominante est inversible.*

Démonstration. Soit $A=[a_{ij}]$ une matrice carrée de taille n à diagonale strictement dominante et soit X un vecteur-colonne non nul, de coefficients $x_1, x_2, \ldots, x_n$. Choisissons un coefficient x_q de module maximum : $|x_q| \geqslant |x_j|$ pour tout j. Puisque X n'est pas le vecteur nul, x_q est non

nul. Le coefficient d'indice q du vecteur $Y = AX$ est $y_q = \sum_{j=1}^n a_{qj}x_j = a_{qq}x_q + \sum_{j\neq q} a_{qj}x_j$. Puisque le module d'une somme est inférieur ou égal à la somme des modules, on a $\left|\sum_{j\neq q} a_{qj}x_j\right| \leqslant \sum_{j\neq q} |a_{qj}||x_j| \leqslant |x_q| \sum_{j\neq q} |a_{qj}| < |x_q||a_{qq}|$ et la dernière inégalité est stricte car A est à diagonale strictement dominante et $|x_q| \neq 0$. On en déduit que y_q n'est pas nul. Ainsi, pour tout vecteur $X \neq 0$, on a $AX \neq 0$: la matrice A est donc inversible. ■

Proposition. *La méthode de relaxation converge dans chacun des cas suivants :*

a) *la matrice A est à diagonale strictement dominante et $0 < \omega < 1$;*

b) *la matrice A est symétrique définie positive et $0 < \omega < 2$.*

Démonstration. On a $(I_n - \omega L)\mathscr{L}_\omega = (1-\omega)I_n + \omega U$, donc

$$(I_n - \omega L)(\mathscr{L}_\omega - zI_n) = (1-\omega-z)I_n + \omega U + z\omega L$$

et comme $I_n - \omega L$ a pour déterminant 1, le polynôme caractéristique de $\mathscr{L}_\omega$ est

$$P(z) = \det(\mathscr{L}_\omega - zI_n) = \det\big[(1-\omega-z)I_n + \omega U + z\omega L\big]$$

Supposons $0<\omega<1$ et soit z un nombre complexe tel que $|z| \geqslant 1$. Puisque $0 < 1-\omega < 1$, on a $1-\omega-z \neq 0$, donc $P(z) = (1-\omega-z)^n \det(I_n + bU + aL)$, où l'on a posé $a = \dfrac{z\omega}{1-\omega-z}$ et $b = \dfrac{\omega}{1-\omega-z}$. D'autre part, l'inégalité $|z|(1-\omega) \geqslant 1-\omega$ s'écrit $|z|\omega \leqslant |z| - (1-\omega)$ et comme on a $|1-\omega-z| \geqslant |z|-(1-\omega) > 0$ d'après l'inégalité triangulaire, on en déduit $|a| = \dfrac{|z\omega|}{|1-\omega-z|} \leqslant 1$. Puisque $\omega \leqslant |z\omega|$, il vient $|b| \leqslant |a| \leqslant 1$. Supposons que A est à diagonale strictement dominante. Alors $I_n + U + L$ est à diagonale strictement dominante (voir les matrices U et L page 250) ; la matrice $I_n + bU + aL$ s'obtient en multipliant les coefficients non diagonaux par des nombres de module au plus 1, donc $I_n + bU + aL$ est aussi à diagonale strictement dominante : ainsi cette matrice est inversible, donc de déterminant non nul. Finalement, si $|z| \geqslant 1$, alors $P(z) \neq 0$. Cela montre que les valeurs propres de la matrice $\mathscr{L}_\omega$ sont de module strictement inférieur à 1, donc la méthode de relaxation converge.

Supposons maintenant A symétrique définie positive et $0 < \omega < 2$. Posons pour simplifier $B = \mathscr{L}_\omega$ et $M = \dfrac{1}{\omega}D - E$. La matrice triangulaire M est inversible et un calcul simple montre que l'on a $M(I_n - B) = A$, ou encore $B = I_n - M^{-1}A$. En utilisant l'égalité ${}^tA = A$, on vérifie en outre la relation suivante :

$$(*) \qquad A - ({}^tB)AB = (I_n - {}^tB)(M + {}^tM - A)(I_n - B)$$

Puisque A est symétrique, on a ${}^tE = F$, ${}^tM = \dfrac{1}{\omega}D - F$ et $M + {}^tM - A = \dfrac{2}{\omega}D - E - F - A = \dfrac{2-\omega}{\omega}D$, car $A + E + F = D$. Les coefficients de D sont les produits $({}^t\boldsymbol{E_i})A\boldsymbol{E_i}$, donc ils sont strictement positifs puisque A est définie positive. Il s'ensuit que la matrice diagonale $\Delta = \dfrac{2-\omega}{\omega}D$ est définie positive, car on a $0<\omega<2$. Soit λ une valeur propre de B et X un vecteur propre associé (puisque A est symétrique, λ est un nombre réel). On a $BX = \lambda X$ et $(I_n - B)X = (1-\lambda)X$. En multipliant les différents termes de $(*)$ à gauche par ${}^tX$ et à droite par X, on obtient $({}^tX)({}^tB)ABX = \lambda^2({}^tX)AX$ et $({}^tX)(I_n - {}^tB)\Delta(I_n - B)X = (1-\lambda)^2({}^tX)\Delta X$, d'où

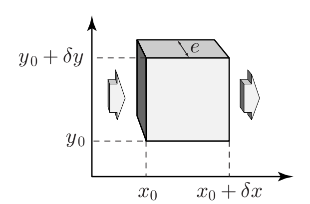
$$(1-\lambda^2)({}^tX)AX = (1-\lambda)^2({}^tX)\Delta X\,.$$

Remarquons que λ ne peut pas être égal à 1 : sinon on aurait $X = BX = X - M^{-1}AX$, donc $M^{-1}AX = 0$, ce qui n'est pas possible car X est non nul et les matrices M^{-1} et A sont inversibles. Ainsi on a $(1-\lambda)^2 > 0$, $({}^tX)\Delta X > 0$ et $({}^tX)AX > 0$, donc aussi $1 - \lambda^2 > 0$, c'est-à-dire $-1 < \lambda < 1$. ■

2.3 Résolution numérique d'une équation de Poisson

Le problème. On considère une plaque en équilibre thermique : en tout point, la température u n'est fonction que des coordonnées (x,y) de ce point. La plaque est convenablement isolée sur ses deux faces de sorte que la conduction s'exerce dans son plan. Le flux de chaleur est proportionnel à la section traversée, au gradient de température $\left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)$ et à la conductivité c, qu'on suppose uniforme.

Exprimons l'équilibre thermique d'un petit volume découpé dans la plaque.



- La face en x_0 a pour surface $e\,\delta y$, où e est l'épaisseur (constante) de la plaque. Le flux rentrant par cette face est donc $-ce\,\delta y \frac{\partial u}{\partial x}(x_0, y_0)$, le signe moins venant de ce que le flux de chaleur se fait dans le sens des températures décroissantes.
- Le flux sortant par la face en $x_0 + \delta x$ est de même $-ce\,\delta y \frac{\partial u}{\partial x}(x_0 + \delta x, y_0)$.

En négligeant $(\delta x)^2$, il vient $\frac{\partial u}{\partial x}(x_0 + \delta x, y_0) = \frac{\partial u}{\partial x}(x_0, y_0) + \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)\delta x$. Le flux qui traverse le volume dans la direction x est donc

$$-ce\,\delta y \left[\frac{\partial u}{\partial x} - \left(\frac{\partial u}{\partial x} + \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)\delta x\right)\right] = ce\,\delta x \delta y \frac{\partial^2 u}{\partial x^2}$$

la dérivée seconde étant prise en (x_0, y_0). De même, dans la direction y, le flux est $ce\,\delta y \delta x \frac{\partial^2 u}{\partial y^2}$. Puisque l'équilibre thermique est supposé atteint, le flux total à travers l'élément est nul :

$$ce\,\delta x \delta y \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = 0$$

d'où l'*équation de Poisson*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0\,.$$

Pour simplifier, nous avons supposé idéalement que la plaque n'échangeait pas de chaleur avec l'extérieur. Dans le cas général où l'on a une distribution de chaleur $f(x, y)$ en tout point de la plaque (réalisée par exemple en insérant un circuit chauffant ou réfrigérant), l'état d'équilibre se traduit par l'égalité

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f$$

et le problème est de calculer la fonction u connaissant son *laplacien* $\Delta u = f$; pour déterminer la solution, il faut imposer les valeurs de u aux bords de la plaque.

Discrétisation. Utilisons l'approximation de la dérivée seconde présentée page 141 : pour une fonction numérique $(x, y) \mapsto u(x, y)$, le nombre

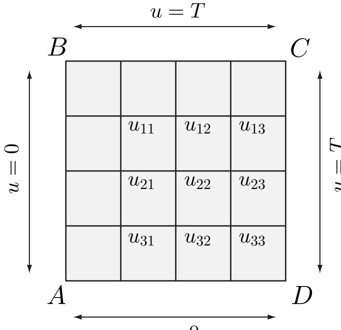
$$\frac{1}{h^2}\left[u(x-h,y)-2u(x,y)+u(x+h,y)\right]$$

est une bonne approximation de $\dfrac{\partial^2 u}{\partial x^2}(x, y)$ si h est assez petit. En tout point (x, y), on peut donc approcher $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}$ par

$$(*) \quad \frac{1}{h^2}\left[u(x-h,y)-2u(x,y)+u(x+h,y)\right]+\frac{1}{k^2}\left[u(x,y-k)-2u(x,y)+u(x,y+k)\right]$$

à la condition que h et k soient assez petits.

Supposons que la plaque est un carré $ABCD$ de côté ℓ et plaçons-nous dans des conditions théoriques où l'on maintient une température constante de zéro degré sur les côtés AB et AD, et une température T sur BC et CD, sauf aux points B et D où la température n'est pas définie. Subdivisons la plaque en seize carrés de côtés $\ell/4$, ce qui fait apparaître neuf points intérieurs qu'on numérote comme sur la figure par les couples (i, j), où $1 \leqslant i \leqslant 3$ et $1 \leqslant j \leqslant 3$. Nous allons estimer la température u_{ij} en chacun de ces points, sous l'hypothèse d'un état stationnaire.



Les solutions doivent *a priori* présenter une symétrie par rapport à la diagonale AC.

Au point $(1, 1)$, $\quad \dfrac{\partial^2 u}{\partial x^2}$ est approché par $(1/h^2)(0 - 2u_{11} + u_{12})$,

$\dfrac{\partial^2 u}{\partial y^2}$ est approché par $(1/k^2)(u_{21} - 2u_{11} + T)$.

Puisqu'on a ici $h = k = \ell/4$, l'égalité de Poisson $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0$ donne

$$0 + T + u_{12} + u_{21} - 4u_{11} = 0\,.$$

Procédons de même pour les autres points (i, j) ; en les ordonnant par balayage des lignes, on obtient le système linéaire carré :

$$\left\{\begin{array}{rcl}
-4u_{11} + u_{12} + u_{21} & = & -T \\
u_{11} - 4u_{12} + u_{13} + u_{22} & = & -T \\
u_{12} - 4u_{13} + u_{23} & = & -2T \\
u_{11} - 4u_{21} + u_{22} + u_{31} & = & 0 \\
u_{12} + u_{21} - 4u_{22} + u_{23} + u_{32} & = & 0 \\
u_{13} + u_{22} - 4u_{23} + u_{33} & = & -T \\
u_{21} - 4u_{31} + u_{32} & = & 0 \\
u_{22} + u_{31} - 4u_{32} + u_{33} & = & 0 \\
u_{23} + u_{32} - 4u_{33} & = & -T
\end{array}\right.$$

de matrice

$$S=\begin{bmatrix}-4&1&0&1&0&0&0&0&0\\1&-4&1&0&1&0&0&0&0\\0&1&-4&0&0&1&0&0&0\\1&0&0&-4&1&0&1&0&0\\0&1&0&1&-4&1&0&1&0\\0&0&1&0&1&-4&0&0&1\\0&0&0&1&0&0&-4&1&0\\0&0&0&0&1&0&1&-4&1\\0&0&0&0&0&1&0&1&-4\end{bmatrix}=\begin{bmatrix}M&I_3&0\\I_3&M&I_3\\0&I_3&M\end{bmatrix}\ ,\ \text{où } M=\begin{bmatrix}-4&1&0\\1&-4&1\\0&1&-4\end{bmatrix}$$

et où 0 représente la matrice carrée nulle de taille 3. Le premier membre du système s'écrit aussi $\begin{bmatrix}M&I_3&0\\I_3&M&I_3\\0&I_3&M\end{bmatrix}\begin{bmatrix}U_1\\U_2\\U_3\end{bmatrix}$, en posant $U_1=\begin{bmatrix}u_{11}\\u_{12}\\u_{13}\end{bmatrix}$, $U_2=\begin{bmatrix}u_{21}\\u_{22}\\u_{23}\end{bmatrix}$ et $U_3=\begin{bmatrix}u_{31}\\u_{32}\\u_{33}\end{bmatrix}$.

Remarquons que la taille de la plaque et ses propriétés thermiques n'interviennent pas dans le système (invariance d'échelle).
La matrice S est à diagonale strictement dominante ; elle est aussi symétrique et comme les coefficients diagonaux de $-S$ sont strictement positifs, $-S$ est définie positive (exercice 3). Observons que la matrice M est tridiagonale.
Avec la méthode de relaxation pour $\omega=0{,}8$ et en prenant le vecteur initial dont toutes les coordonnées valent $0{,}5T$, on obtient en quatre itérations la solution

$$u_{11}=u_{22}=u_{33}=0{,}5T\ ,\ u_{12}=u_{23}=0{,}714T\ ,\ u_{21}=u_{32}=0{,}285T\ ,\ u_{13}=0{,}857T\ ,\ u_{31}=0{,}142T$$

avec une erreur relative inférieure à un centième.

Généralité de la méthode. Dans les applications, on utilise une subdivision beaucoup plus fine et la matrice du système est de grande taille. Supposons qu'il y a n^2 points dans la subdivision ; ordonnons-les en parcourant ligne par ligne (ou colonne par colonne) ; en utilisant l'approximation $(*)$ du laplacien, on obtient la matrice carrée de taille n^2

$$S=\begin{bmatrix}M&I_n&&&\\I_n&M&I_n&0&\\&\ddots&\ddots&\ddots&\\&0&I_n&M&I_n\\&&&I_n&M\end{bmatrix}\ ,\ \text{où } M=\begin{bmatrix}b&a&&&\\a&b&a&0&\\&\ddots&\ddots&\ddots&\\&0&a&b&a\\&&&a&b\end{bmatrix},\ \ a=\frac{1}{h^2},\ b=-\frac{2}{h^2}-\frac{2}{k^2}.$$

La matrice S est toujours symétrique à diagonale strictement dominante et $-S$ est définie positive. La méthode de relaxation est donc appropriée pour résoudre numériquement une équation de Poisson.

Quand le domaine est moins régulier, on le subdivise par des triangles pour mieux épouser son bord.

Remarque

De nombreuses équations aux dérivées partielles font intervenir, comme l'équation

de Poisson, le laplacien de u défini, dans le cas d'une fonction de deux variables cartésiennes, par $\Delta u = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}$. L'opération $u \mapsto \Delta u$ est linéaire.

Par exemple, pour étudier les vibrations d'une membrane ou la surface d'un liquide en mouvement dans un bassin, on est amené à résoudre l'équation $\Delta u = -k^2 u$: les solutions u sont des vecteurs propres de l'opérateur laplacien (voir l'exercice 11 page 573). En utilisant la méthode de discrétisation précédente, on est conduit au système linéaire $SU = -k^2 U$ et à calculer les valeurs propres de S.

3. Calcul de valeurs propres

Si A est une matrice carrée de grande taille, calculer son déterminant est très coûteux en nombre d'opérations : passer par le polynôme caractéristique n'est donc pas toujours numériquement efficace pour trouver les valeurs propres. De plus, on n'obtient en général qu'une valeur approchée $\tilde{\lambda}$ de la valeur propre, de sorte que l'équation linéaire $AX = \tilde{\lambda} X$ n'a pour solution que le vecteur nul : utiliser ce système pour le calcul d'un vecteur propre présente de sérieuses difficultés. Voici une méthode itérative qui, lorsqu'elle converge, produit des vecteurs X_k ayant pour limite un vecteur propre de A.

Un exemple. Soit la matrice $A = \begin{bmatrix} 1 & 7 \\ 0 & 10 \end{bmatrix}$, de valeurs propres 1 et 10. Posons $Y_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $X_0 = \dfrac{1}{\|Y_0\|} Y_0$ et appliquons l'algorithme suivant :

$$\text{pour } k \text{ de } 1 \text{ à } p, \text{ faire} \quad \begin{cases} (i) & Y_k \leftarrow AX_{k-1} \\ (ii) & X_k \leftarrow \dfrac{1}{\|Y_k\|} Y_k \end{cases}$$

Pour $p = 4$, on obtient le vecteur X_4 de coordonnées $(0{,}6139\ldots, 0{,}7893\ldots)$ et les coordonnées de AX_4 sont $(6{,}139\ldots, 7{,}893\ldots)$; ainsi AX_4 est peu différent de $10X_4$, autrement dit le vecteur X_4 est à peu près propre pour la valeur propre 10. Pour des valeurs de p plus grandes, la précision s'améliore.

Explication. Les vecteurs X_k sont obtenus en itérant X_0 par la transformation $W \mapsto AW$ et en divisant chaque fois par la norme. On a $A^k = \begin{bmatrix} 1 & a_k \\ 0 & 10^k \end{bmatrix}$, avec $a_k = 7\dfrac{10^k - 1}{9}$, et il vient $W_k = A^k X_0 = \dfrac{1}{\sqrt{5}} A^k Y_0 = \dfrac{1}{\sqrt{5}} \begin{bmatrix} 1 + 2a_k \\ 2.10^k \end{bmatrix}$. Posons $\begin{bmatrix} u_k \\ v_k \end{bmatrix} = W_k$. Alors $\dfrac{v_k}{v_{k-1}} = 10$ et comme a_k tend vers l'infini, on a $\lim \dfrac{u_k}{u_{k-1}} = \lim \dfrac{a_k}{a_{k-1}} = \lim \dfrac{10^k - 1}{10^{k-1} - 1} = 10$. Ainsi le rapport des coefficients entre W_k et W_{k-1} tend vers 10 quand k tend vers l'infini. Puisque $X_k = \dfrac{W_k}{\|W_k\|}$, on passe aussi à peu près de X_{k-1} à X_k en multipliant par 10, lorsque k est assez grand. Si les X_k tendent vers un vecteur X, on a donc $AX = 10X$, autrement dit X est un vecteur propre pour la valeur propre 10.

En calculant les formules générales pour les coefficients de X_k, on voit facilement que $\lim X_k$ est colinéaire au vecteur $V = \begin{bmatrix} 7 \\ 9 \end{bmatrix}$ qui est propre pour la valeur propre 10 (exercice 5).

Hypothèses

Dans la suite, A est une matrice carrée non nulle de taille n à coefficients réels (pour une matrice complexe, voir la fin de la remarque page 257).

a) On suppose que l'une des valeurs propres de A, disons λ, est de module strictement supérieur aux autres et que λ est valeur propre simple.

b) X_0 est un vecteur ayant, dans une base de trigonalisation, une coordonnée non nulle relativement au vecteur propre associé à λ.

L'hypothèse (a) implique que λ est réel : en effet, puisque A est à coefficients réels, le nombre conjugué $\bar{\lambda}$ est une valeur propre de même module que λ, donc $\bar{\lambda}$ n'est pas différent de λ.

Proposition. *Définissons les vecteurs Y_k et X_k en posant $Y_k = AX_{k-1}$ et $X_k = \frac{1}{\|Y_k\|} Y_k$ pour $k \geqslant 1$. Alors*

i) *les nombres $\|Y_k\|$ tendent vers $|\lambda|$,*

ii) *les vecteurs $\varepsilon^k X_k$ tendent vers un vecteur propre pour λ, où $\varepsilon = 1$ si $\lambda > 0$ et $\varepsilon = -1$ si $\lambda < 0$.*

Démonstration. Traitons d'abord le cas d'une matrice de la forme $B = \begin{bmatrix} \lambda & 0 \\ 0 & T \end{bmatrix}$, où T est de taille $n-1$ et triangulaire ; les coefficients diagonaux $\lambda_2, \ldots, \lambda_n$ de T sont des valeurs propres de B autres que λ. Par hypothèse, on a $|\lambda| > |\lambda_i|$, donc λ n'est pas nul et les coefficients diagonaux de la matrice $\lambda^{-1}T$ sont de module strictement inférieur à 1 : il s'ensuit que $(\lambda^{-1}T)^k$ tend vers 0 quand k tend vers l'infini (proposition page 185). Écrivons les vecteurs sous la forme $\begin{bmatrix} u \\ V \end{bmatrix}$, où u est un nombre et V un vecteur à $n-1$ coordonnées. Si $U_0 = \begin{bmatrix} u_0 \\ V_0 \end{bmatrix}$, les vecteurs $U_k = B^k U_0$ sont donnés par $U_k = \begin{bmatrix} \lambda^k & 0 \\ 0 & T^k \end{bmatrix} \begin{bmatrix} u_0 \\ V_0 \end{bmatrix} = \begin{bmatrix} \lambda^k u_0 \\ T^k V_0 \end{bmatrix} = \lambda^k U'_k$, avec $U'_k = \begin{bmatrix} u_0 \\ (\lambda^{-1}T)^k V_0 \end{bmatrix}$. En utilisant une norme $\| \ \|_*$ quelconque, il vient $\|U_k\|_* = |\lambda|^k \|U'_k\|_*$. Les vecteurs U'_k tendent vers $\begin{bmatrix} u_0 \\ 0 \end{bmatrix} = u_0 \boldsymbol{E_1}$ et $\|U'_k\|_*$ tend vers $\|u_0 \boldsymbol{E_1}\|_* = |u_0| \|\boldsymbol{E_1}\|_*$, car une norme est continue. En posant $Z_k = \frac{1}{\|U_k\|_*} U_k = \frac{\lambda^k}{|\lambda|^k} \frac{1}{\|U'_k\|_*} U'_k$ et $\varepsilon = \frac{\lambda}{|\lambda|}$, on voit que $\varepsilon^k Z_k = \frac{1}{\|U'_k\|_*} U'_k$ tend vers $\frac{u_0}{|u_0| \|\boldsymbol{E_1}\|_*} \boldsymbol{E_1}$ pourvu que u_0 ne soit pas nul. Puisque $\boldsymbol{E_1}$ est un vecteur propre associé à λ, la propriété (ii) est vraie dans le cas de notre matrice B particulière et pour une norme quelconque. De plus, $BZ_k = \frac{1}{\|U_k\|_*} U_{k+1} = \frac{1}{|\lambda|^k \|U'_k\|_*} \lambda^{k+1} U'_{k+1}$ et $\|BZ_k\|_* = |\lambda| \frac{\|U'_{k+1}\|_*}{\|U'_k\|_*}$ tend vers $|\lambda|$, d'où (i).

Dans le cas général, il existe une base de $\mathbb{C}^n$ dans laquelle la transformation $f : X \mapsto AX$ a une matrice B de la forme précédente (page 182) et si P est la matrice de passage, alors $B^k = P^{-1} A^k P$ pour tout entier $k \geqslant 0$. Soient $W_0 \in \mathbb{R}^n$ et U_0 tel que $W_0 = PU_0$. Les itérés $W_k = A^k W_0$ et les itérés $U_k = B^k U_0$ sont reliés par la relation $W_k = PU_k$. Un vecteur propre associé à λ est la première colonne de P : si l'on suppose que W_0 a une coordonnée non

nulle sur ce vecteur propre, alors la première coordonnée de U_0 est non nulle. Appliquons la première partie de la démonstration en choisissant la norme $\|X\|_* = \|PX\|$ qui provient du produit scalaire $(PX') \cdot (PY')$. Les vecteurs $\varepsilon^k \dfrac{1}{\|U_k\|_*} U_k$ tendent vers un vecteur $\alpha \boldsymbol{E_1}$ (où $\alpha \neq 0$), donc les vecteurs $X_k = \dfrac{1}{\|W_k\|} W_k = \dfrac{1}{\|PU_k\|} PU_k = \dfrac{1}{\|U_k\|_*} PU_k = P\left[\dfrac{1}{\|U_k\|_*} U_k\right]$ tendent vers $\alpha P \boldsymbol{E_1}$: ce vecteur est colinéaire à la première colonne de P, donc c'est un vecteur propre de f pour la valeur propre λ. On a $AX_k = \dfrac{1}{\|W_k\|} W_{k+1} = \dfrac{1}{\|PU_k\|} PU_{k+1} = \dfrac{1}{\|U_k\|_*} PU_{k+1} = PBZ_k$ donc $\|AX_k\| = \|BZ_k\|_*$: d'après ce qui précède, $\|AX_k\|$ tend vers $|\lambda|$. ■

Remarques sur la convergence

- Appelons $\lambda, \lambda_2, \ldots, \lambda_q$ les valeurs propres de A, où $|\lambda| > |\lambda_i|$ pour tout i. D'après la démonstration, la convergence est d'autant plus rapide que les rapports $|\lambda|/|\lambda_i|$ sont grands.

  Pour augmenter ces rapports et donc la vitesse de convergence, on peut être amené à appliquer la méthode à la matrice $A' = A - \alpha I_n$, où α est convenablement choisi ; en ajoutant α à une valeur propre de A', on obtient une valeur propre de A (voir exercice 6).

- Quel que soit le signe de λ, les vecteurs X_{2k} tendent vers un vecteur propre associé à λ.
- Si A est une matrice à coefficients complexes vérifiant les hypothèses que nous avons faites pour la proposition, alors (i) reste vrai et si λ a pour argument θ, les vecteurs $(e^{-\boldsymbol{i}k\theta})X_k$ tendent vers un vecteur propre associé à λ.

Calcul de la valeur propre de plus petit module. Soit A une matrice inversible ayant une valeur propre simple μ de plus petit module. Aucune valeur propre de A n'est nulle, les valeurs propres de A^{-1} sont les inverses des valeurs propres de A, donc $1/\mu$ est la valeur propre de plus grand module de A^{-1}. On peut ainsi calculer une approximation de μ en appliquant la méthode à la matrice A^{-1}.

Exemple. Soit la matrice symétrique $A = \dfrac{1}{150}\begin{bmatrix} 172 & -161 & -154 & -237 \\ -161 & 743 & 477 & 256 \\ -154 & 477 & 368 & 79 \\ -237 & 256 & 79 & 37 \end{bmatrix}$. En prenant pour X_0 le vecteur dont toutes les coordonnées valent 1, on obtient $\|Y_5\| = 7{,}999996\ldots$ et les coordonnées de X_5 sont à peu près $(-0{,}258, 0{,}774, 0{,}516, 0{,}258)$: en fait, la plus grande valeur propre est $\lambda = 8$ et $V = (-1, 3, 2, 1)$ est un vecteur propre associé ; les deux vecteurs $(1/\|V\|)V$ et X_5 sont de norme 1 et leur écart est de norme inférieure à $1{,}6\,10^{-4}$.
En appliquant la méthode à A^{-1}, on trouve que la valeur propre de A de plus petite valeur absolue est $\mu = 1/2$, avec vecteur propre $W = (1, 2, -3, 1)$.
Puisque la matrice A est symétrique, les vecteurs propres relatifs à des valeurs propres différentes de λ et μ sont dans le plan orthogonal aux vecteurs V et W (énoncé page 216). Choisissons deux vecteurs formant une base de ce plan, par

exemple $U = \begin{bmatrix} 13 \\ 1 \\ 5 \\ 0 \end{bmatrix}$ et $U' = \begin{bmatrix} 13 \\ -9 \\ 7 \\ 26 \end{bmatrix}$. Si P est la matrice de colonnes V, W, U, U', alors on a $P^{-1}AP = \begin{bmatrix} D & 0 \\ 0 & M \end{bmatrix}$, où $D = \begin{bmatrix} 8 & 0 \\ 0 & 1/2 \end{bmatrix}$ et $M = \frac{1}{130} \begin{bmatrix} 168 & -108 \\ -81 & -129 \end{bmatrix}$. Les autres valeurs propres de A sont celles de la matrice M ; elles sont faciles à calculer puisque le polynôme caractéristique de M est de degré 2 : on trouve $3/2$ et $-6/5$.

Exercices

@ **1. Conditionnement d'une matrice symétrique définie positive.** Soit S une matrice symétrique définie positive. Montrer que le conditionnement de S est égal à $v_{\max}/v_{\min}$, où $v_{\max}$ est la plus grande valeur propre de S et $v_{\min}$ la plus petite (appliquer la proposition page 245).

@ **2. Localisation des valeurs propres.** Soit $A = [a_{ij}]$ une matrice carrée de taille n et soit $\lambda \in \mathbb{C}$ une valeur propre de A.

a) Soit $u = (u_1, u_2, \ldots, u_n)$ un vecteur propre relatif à λ et soit k un indice tel que $|u_k| \geqslant |u_i|$ pour tout i. Montrer que l'on a $|\lambda - a_{kk}| \leqslant \sum_{j \neq k} |a_{kj}|$.

b) Si a est un nombre complexe et r un nombre réel positif, posons $D(a, r) = \{z \in \mathbb{C} \mid |z - a| \leqslant r\}$: c'est le disque de rayon r centré au point d'affixe a. Montrer que les valeurs propres de A sont dans la réunion des disques $D\big(a_{pp}, \sum_{j \neq p} |a_{pj}|\big)$.

c) En déduire l'inégalité $|\lambda| \leqslant \max_k \left(\sum_{j=1}^{n} |a_{kj}| \right)$.

@ **3. Des matrices sympathiques.** Soit $A = [a_{ij}]$ une matrice carrée de taille n à diagonale strictement dominante et à coefficient diagonaux tous strictement positifs.

a) Soit λ une valeur propre complexe de A. En utilisant l'inégalité obtenue en **(a)** de l'exercice précédent, montrer que la partie réelle de λ est strictement positive.

b) Montrer que si de plus A est symétrique et à coefficients réels, alors A est définie positive.

4. Montrer qu'une matrice carrée A à diagonale strictement dominante possède une décomposition LU (remarquer que les matrices $A^{(k)}$ sont aussi à diagonale strictement dominante et appliquer une proposition page 250).

5. Reprenons la matrice $A = \begin{bmatrix} 1 & 7 \\ 0 & 10 \end{bmatrix}$ de l'exemple page 255 et les vecteurs X_k définis par $X_0 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ et $X_{k+1} = \frac{1}{\|AX_k\|} AX_k$. Calculer les coefficients de X_k pour tout $k \geqslant 1$ et la limite des vecteurs X_k.

@ **6. Une méthode itérative pour résoudre certaines équations linéaires.** Soient A une matrice carrée de taille n et B un vecteur-colonne de $\mathbb{K}^n$. On considère le système d'équations linéaires $(S)\ :\ AX = B$.

a) Soit a un nombre non nul. Montrer que X est solution de (S) si et seulement X est point fixe de la transformation affine $T : Z \mapsto (I_n - aA)Z + aB$.

On en déduit une technique pour résoudre (S) : chercher un nombre a tel que la matrice $I_n - aA$ ait toutes ses valeurs propres de module strictement inférieur à 1. S'il en est ainsi, alors pour tout vecteur initial X_0, les itérés $X_1 = T(X_0), \dots, X_p = T(X_{p-1}), \dots$ ont pour limite la solution de (S) (corollaire page 185).

b) Supposons que toutes les valeurs propres de A sont réelles et comprises entre les nombres positifs u et v (où $u < v$). Posons $a = \dfrac{2}{v+u}$. Montrer que les valeurs propres de la matrice $M = I_n - aA$ sont de valeur absolue inférieure à $\dfrac{v-u}{v+u} < 1$ (les valeurs propres de M sont les nombres $1 - \lambda a$, où λ est valeur propre de A).

c) Application. On prend $A = \begin{bmatrix} 3 & 1/2 & 0 & 0 & 0 \\ 1/2 & 3 & 1/3 & 0 & 0 \\ 0 & 1/3 & 3 & 1/3 & 0 \\ 0 & 0 & 1/3 & 3 & 1/2 \\ 0 & 0 & 0 & 1/2 & 3 \end{bmatrix}$.

(i) Montrer que les valeurs propres de A sont réelles et comprises entre $u = 2{,}15$ et $v = 3{,}84$ (utiliser l'exercice **2.b)**). Calculer a.

(ii) Considérons les itérés $X_1 = T(X_0), \dots, X_{p+1} = T(X_p), \dots$ d'un vecteur $X_0 \in \mathbb{R}^5$. Posons $B_p = AX_p$ et notons $\delta_p = \|X - X_p\|$ l'erreur commise en remplaçant la solution X de (S) par son approximation X_p. Montrer que l'on a $\delta_p \leqslant \operatorname{cond}(A)\|B - B_p\|\|X_p\|/\|B_p\|$.

(iii) Montrer que le conditionnement de A est inférieur ou égal à v/u (remarquer que la matrice A est symétrique ; en utilisant **(i)**, montrer qu'elle est définie positive et appliquer l'exercice (1).

(iv) Écrire un algorithme permettant de calculer avec une précision ε donnée, les coordonnées du vecteur X solution de $AX = B$.

(v) Supposons $B = (1, 0, 2, 0, 1)$. Montrer que si l'on prend $X_0 = \mathbf{0}$, les coordonnées de X sont celles de X_4 à 0,002 près.

Chapitre 9

Limites, dérivées, intégrales

1. Rappels sur les limites

1.1 Limite d'une suite

Au chapitre 1, nous avons expliqué pourquoi, dans une suite décroissante de nombres positifs, les décimales de ces nombres se stabilisent (page 6). On a la même propriété pour une suite décroissante et minorée par un nombre quelconque, ou pour une suite croissante majorée. Rappelons quelques définitions :

- Une suite (u_n) est *croissante* si l'on a $u_{n+1} \geqslant u_n$ pour tout n.
- Une suite (u_n) est *majorée* s'il existe un nombre M supérieur ou égal à tous les termes de la suite, c'est-à-dire que l'on a $u_n \leqslant M$ pour tout n.
  Une suite (u_n) est *minorée* s'il existe un nombre m tel que $m \leqslant u_n$ pour tout n.
- Une suite (u_n) est *convergente* si elle a une limite finie ℓ : cela veut dire que pour tout nombre $\varepsilon > 0$, tous les termes u_n sont, après un certain rang, compris entre $\ell - \varepsilon$ et $\ell + \varepsilon$.

Énonçons les propriétés essentielles des suites monotones.

► *Soit (u_n) une suite croissante de nombres réels. Si (u_n) est majorée, elle est convergente ; si (u_n) n'est pas majorée, alors u_n tend vers $+\infty$.*

► *Soit (u_n) une suite décroissante de nombres réels. Si (u_n) est minorée, elle est convergente ; si (u_n) n'est pas minorée, alors u_n tend vers $-\infty$.*

Suites adjacentes

Des suites (a_n) et (b_n) sont dites *adjacentes* si

i) (a_n) est croissante et (b_n) est décroissante,

ii) $a_n \leqslant b_n$ pour tout n,

iii) $b_n - a_n$ tend vers 0.

Dans ce cas, la suite (a_n) est majorée par b_0, la suite (b_n) est minorée par a_0, donc la suite (a_n) a une limite ℓ et la suite (b_n) a une limite ℓ' ; d'après les théorèmes sur les limites, la suite $(b_n - a_n)$ a pour limite $\ell' - \ell$ et comme par hypothèse $b_n - a_n$ tend vers 0, on en déduit $\ell = \ell'$; la suite (a_n) étant croissante, ℓ est supérieur ou égal à tous les a_n et de même, ℓ' est inférieur ou égal à tous les b_p.

Des suites adjacentes (a_n) et (b_n) ont la même limite ℓ et l'on a $a_n \leqslant \ell \leqslant b_p$ pour tous n et p.

Puisque l'écart entre a_n et b_n tend vers 0, ces nombres permettent d'encadrer la limite ℓ avec la précision qu'on veut.

Suites itératives

Nous avons déjà rencontré des suites (u_n) formées des itérés d'un nombre initial u_0 par une transformation f : on a par définition

$$u_{n+1} = f(u_n), \text{ pour tout } n \geqslant 0.$$

Faisons deux hypothèses sur la fonction f.

Hypothèse (a) : f a un point fixe p, c'est-à-dire tel que $f(p) = p$.

Pour visualiser le comportement des u_n, on peut procéder comme page 15 : sur un même dessin, représentons le graphe de $y=f(x)$ et la droite d'équation $y=x$; ces deux graphes se coupent donc au point $A = (p, p)$. Construisons une ligne brisée comme ci-dessous, en joignant successivement les points (u_0,u_1), (u_1,u_1), (u_1,u_2), (u_2,u_2), etc :

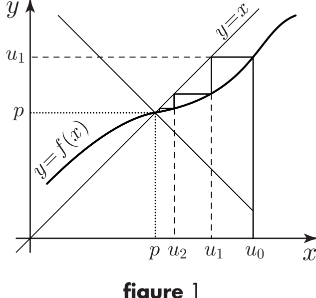


figure 1

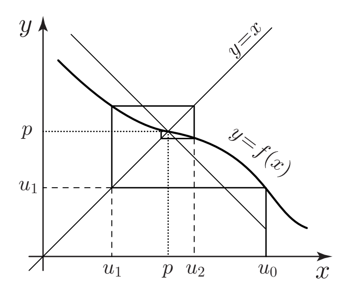


figure 2

Hypothèse (b) : Au voisinage du point A, le graphe de f reste dans le cône (partie grisée de la figure 3) formé de deux droites sécantes en A de pente $\pm k$, où k est un nombre tel que $0 < k < 1$.

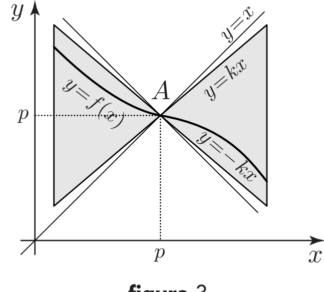


figure 3

Si $M = (x, y)$ est un point du cône, la pente du segment AM est comprise entre $-k$ et k, autrement dit $|y - p| \leqslant k|x - p|$.

Si x est l'abscisse d'un point du cône, alors le point $\big(x, f(x)\big)$ est dans le cône et donc $\big|f(x) - p\big| \leqslant k|x - p|$. Puisque $k < 1$, il en résulte que $f(x)$ est plus proche de p que x ; en particulier, $f(x)$ est aussi l'abscisse d'un point du cône.

Définition
Si les hypothèses (a) et (b) sont vérifiées, on dit que f est *contractante au voisinage du point fixe* A. Le nombre $k < 1$ s'appelle *le coefficient de contraction*.

Si la valeur initiale u_0 de la suite est l'abscisse d'un point du cône, alors le point $M_0 = (u_0, u_1)$ est dans le cône, donc aussi le point $M_1 = (u_1, u_2)$ et de proche en proche, tous les points $M_n = (u_n, u_{n+1})$ sont dans le cône. On a donc

$$|u_{n+1} - p| \leqslant k|u_n - p|, \quad \text{pour tout } n \geqslant 0$$

Il s'ensuit

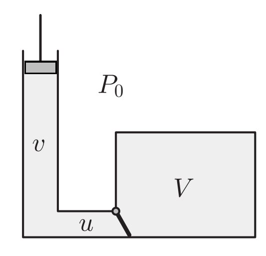
$$(*) \qquad |u_n - p| \leqslant k|u_{n-1} - p| \leqslant k^2|u_{n-2} - p| \leqslant \cdots \leqslant k^n|u_0 - p|$$

Puisqu'on a supposé $0 < k < 1$, k^n tend vers 0 quand n tend vers l'infini, donc $|u_n - p|$ tend vers 0, autrement dit u_n tend vers p.

> *Si f est contractante au voisinage du point fixe p et si u_0 est assez proche de p, alors la suite définie par $u_{n+1} = f(u_n)$ a pour limite p.*

On dit que le point fixe p est *attractif*. Les inégalités $(*)$ montrent que la convergence est d'autant plus rapide que le coefficient de contraction k est plus petit.

Exemple. Une pompe de volume v permet d'envoyer de l'air dans une enceinte de volume V par un tuyau d'alimentation de petit volume u. On suppose que la compression est isotherme. Quelle est la pression dans l'enceinte après n coups de pompe ?



La pompe est à la pression extérieure P_0 qui est aussi la pression initiale dans l'enceinte. Supposons qu'on a obtenu une pression P dans l'enceinte et donnons un coup de pompe. La valve s'ouvre lorsque la pompe exerce une pression P, le volume d'air dans celle-ci étant passé de v à une valeur $v_1 < v$. Le produit de la pression par le volume est constant, donc $P_0(v+u) = P(v_1+u)$. Quand le piston arrive en fin de course, le volume v_1+u+V a été comprimé en $u+V$ et la pression dans l'enceinte et le tuyau prend la valeur P' telle que $P(v_1 + u + V) = P'(V + u)$. Ainsi on a $P'(V + u) = P(v_1 + u) + PV = P_0(v + u) + PV$. En notant P_n la pression dans l'enceinte après n coups de pompe, il vient la relation

$$P_{n+1} = \frac{V}{V+u} P_n + \frac{v+u}{V+u} P_0$$

► La fonction affine $f : x \mapsto \dfrac{V}{V+u} x + \dfrac{v+u}{V+u} P_0$ a un point fixe p :

$$p = \frac{V}{V+u} p + \frac{v+u}{V+u} P_0, \quad \text{d'où} \quad p = \frac{v+u}{u} P_0.$$

► Le graphe de f est la droite de pente $k = \dfrac{V}{V+u} < 1$ passant par le point (p, p) : l'hypothèse (b) est donc satisfaite.

On en déduit que P_n tend vers la pression limite $\frac{v+u}{u}P_0$, valeur d'autant plus grande que le volume u du raccord est petit.

1.2 Limite d'une fonction

On a les mêmes propriétés que pour les suites. Rappelons qu'une fonction f à valeurs réelles est *majorée* s'il existe un nombre M tel que $f(x) \leqslant M$ pour tout x. Une fonction f est *minorée* s'il existe un nombre m tel que $f(x) \geqslant m$ pour tout x.
Soit f une fonction définie sur un intervalle $]a,b[$; a et b sont des nombres ou bien l'un des symboles $-\infty$ ou $+\infty$.

- *Si f est croissante et majorée, alors $f(x)$ a une limite finie quand x tend vers b ; si f est croissante et non majorée, alors $\lim\limits_{x\to b} f(x) = +\infty$.*
- *Si f est croissante et minorée, alors $f(x)$ a une limite finie quand x tend vers a ; si f est croissante et non minorée, alors $\lim\limits_{x\to a} f(x) = -\infty$.*

Une fonction décroissante minorée a une limite finie quand x tend vers b ; si f est décroissante non minorée, alors $f(x)$ tend vers $-\infty$ quand x tend vers b.
Voici les propriétés générales du calcul des limites ; nous les formulons pour les fonctions, mais elles sont valables aussi bien pour des suites de nombres. x_0 désigne un nombre ou bien l'un des symboles $+\infty$ ou $-\infty$.
Supposons que $f(x)$ et $g(x)$ ont une limite finie quand x tend vers x_0.

- $\lim\limits_{x\to x_0}\big[f(x)+g(x)\big] = \lim\limits_{x\to x_0} f(x) + \lim\limits_{x\to x_0} g(x)$ *et* $\lim\limits_{x\to x_0}\big[f(x)g(x)\big] = \big[\lim\limits_{x\to x_0} f(x)\big]\big[\lim\limits_{x\to x_0} g(x)\big]$
- *Si* $\lim\limits_{x\to x_0} g(x) \neq 0$, *alors* $\lim\limits_{x\to x_0} \dfrac{f(x)}{g(x)} = \dfrac{\lim\limits_{x\to x_0} f(x)}{\lim\limits_{x\to x_0} g(x)}$.
- *Si $f(x)$ tend vers $+\infty$ ou vers $-\infty$ quand x tend vers x_0, alors* $\lim\limits_{x\to x_0} \dfrac{1}{f(x)} = 0$.

2. Ordres de grandeur

On a souvent besoin de comparer les ordres de grandeur de deux quantités qui tendent simultanément vers 0 ou vers l'infini. Dans ce but, nous allons définir la notion de quantité négligeable devant une autre et de quantités équivalentes.
Chacune des fonctions x, x^2 ou x^n tend vers 0 quand x tend vers 0, mais avec des ordres de grandeur différents : par exemple, pour x proche de 0, x^{n+1} est très petit devant x^n puisque $x^{n+1}/x^n = x$ tend vers 0. Ces fonctions puissances servent d'échelle de comparaison pour mesurer l'ordre de grandeur d'une quantité qui tend vers 0.

2.1 Exemples

Posons $f(x) = \tan x - \sin x$. Puisque $\sin x$ et $\tan x$ tendent vers 0 quand x tend vers 0, il en va de même de leur différence $f(x)$. Pour comparer $f(x)$ aux quantités x, x^2,

x^3 pour de petites valeurs de x, formons les rapports $\frac{f(x)}{x}$, $\frac{f(x)}{x^2}$, $\frac{f(x)}{x^3}$ et calculons leurs valeurs pour $x = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ avec une précision relative de $1/1000$.

x	10^{-1}	10^{-2}	10^{-3}	10^{-4}
$\frac{f(x)}{x}$	$5{,}01.10^{-3}$	5.10^{-5}	5.10^{-7}	5.10^{-9}
$\frac{f(x)}{x^2}$	$5{,}01.10^{-2}$	5.10^{-3}	5.10^{-4}	5.10^{-5}
$\frac{f(x)}{x^3}$	0,501	0,5	0,5	0,5
$\frac{f(x)}{x^4}$	5,01	50	500	5000

Les deux premières lignes du tableau montrent que pour x petit, $f(x)$ est négligeable devant x et devant x^2 ; la troisième ligne indique que $\frac{f(x)}{x^3}$ semble tendre vers $1/2$: cela veut dire que pour x assez petit, $f(x)$ est de l'ordre de grandeur de $(1/2)x^3$: on dit que les quantités $f(x)$ et $(1/2)x^3$ sont des infiniments petits équivalents.
Faisons des calculs analogues pour la différence $g(x) = \tan x - (\sin x)^2$ qui tend aussi vers 0 quand x tend vers 0.

x	10^{-1}	10^{-2}	10^{-3}	10^{-4}
$\frac{g(x)}{x}$	0,903	0,990	0,999	0,9999
$\frac{g(x)}{x^2}$	9,03	99,0	999	9999

Cette fois, $\frac{g(x)}{x}$ semble tendre vers 1, donc $g(x)$ est de l'ordre de x quand x tend vers 0 : on dit que $g(x)$ est un infiniment petit équivalent à x.

2.2 Infiniments petits, infiniments grands

Soient u et v des fonctions définies sur un même intervalle I et soit a un nombre appartenant à I ou bien une extrémité de I (éventuellement $+\infty$ ou $-\infty$)

Définitions

- Si $\frac{u(x)}{v(x)}$ tend vers 0 quand x tend vers a, on dit que $u(x)$ *est infiniment petit*, ou *négligeable*, devant devant $v(x)$ quand x tend vers a, ce que l'on note $u(x) \underset{x\to a}{\ll} v(x)$. On dit aussi que $v(x)$ *est infiniment grand devant* $u(x)$, ce qui se note $v(x) \underset{x\to a}{\gg} u(x)$.
- Si $\frac{u(x)}{v(x)}$ tend vers 1 quand x tend vers a, on dit que $u(x)$ *et* $v(x)$ *sont équivalents quand* x *tend vers* a, ce qui se note $u(x) \underset{x\to a}{\sim} v(x)$.

Propriétés de ces relations

1) *$u(x) \underset{x\to a}{\ll} 1$ si et seulement si $u(x)$ tend vers 0 quand x tend vers a.*

2) *Si $u(x) \underset{x\to a}{\sim} v(x)$ et si $v(x)$ a une limite (éventuellement infinie) quand x tend vers a, alors $u(x)$ a la même limite.*

3) *Si ℓ est un nombre non nul, alors $u(x) \underset{x\to a}{\sim} \ell$ si et seulement si $u(x)$ tend vers ℓ quand x tend vers a.*

4) *Si $u(x) \underset{x\to a}{\sim} v(x)$, alors $u(x)$ et $v(x)$ ont le même signe pour x assez proche de a.*

5) *$u(x) \underset{x\to a}{\sim} v(x)$ si et seulement si $\big[u(x)-v(x)\big] \underset{x\to a}{\ll} v(x)$.*

6) *Si $u(x) \underset{x\to a}{\ll} v(x)$ et si $v(x) \underset{x\to a}{\ll} w(x)$, alors $u(x) \underset{x\to a}{\ll} w(x)$.*

7) *Si $u(x) \underset{x\to a}{\sim} v(x)$ et si $v(x) \underset{x\to a}{\sim} w(x)$, alors $u(x) \underset{x\to a}{\sim} w(x)$.*
Si $u(x) \underset{x\to a}{\sim} v(x)$ et si $v(x) \underset{x\to a}{\ll} w(x)$, alors $u(x) \underset{x\to a}{\ll} w(x)$

8) *Si $u(x) \underset{x\to a}{\sim} v(x)$ et si $w(x) \underset{x\to a}{\sim} h(x)$, alors $u(x)w(x) \underset{x\to a}{\sim} v(x)h(x)$.*
Si $u(x) \underset{x\to a}{\sim} v(x)$ et si $w(x) \underset{x\to a}{\ll} h(x)$, alors $u(x)w(x) \underset{x\to a}{\ll} v(x)h(x)$.

Démonstration. La propriété $u(x) \underset{x\to a}{\ll} 1$ veut dire que $u(x)=\dfrac{u(x)}{1}$ tend vers 0 quand x tend vers a. Supposons $u(x) \underset{x\to a}{\sim} v(x)$. On a $u(x)=\dfrac{u(x)}{v(x)}v(x)$ et $\lim\limits_{x\to a}\dfrac{u(x)}{v(x)}=1$, donc $u(x)$ et $v(x)$ ont la même limite, ce qui montre (2). La propriété (3) vient de ce que si $\ell\neq 0$, $u(x)$ tend vers ℓ si et seulement si $\dfrac{u(x)}{\ell}$ tend vers 1. Si $u(x)$ et $v(x)$ sont équivalents quand x tend vers a, $u(x)/v(x)$ tend vers 1 ; alors pour x assez proche de a, le rapport $u(x)/v(x)$ est positif, donc $u(x)$ et $v(x)$ sont de même signe. Montrons (5) : puisque $\dfrac{u(x)-v(x)}{v(x)}=\dfrac{u(x)}{v(x)}-1$, le rapport $\dfrac{u(x)}{v(x)}$ tend vers 1 si et seulement si $\dfrac{u(x)-v(x)}{v(x)}$ tend vers 0. La propriété (6) affirme simplement que si chacun des rapports $\dfrac{u(x)}{v(x)}$ et $\dfrac{v(x)}{w(x)}$ tend vers 0, leur produit $\dfrac{u(x)}{w(x)}$ tend vers 0. Les quatre dernières propriétés se démontrent de même en utilisant que la limite d'un produit est égale au produit des limites. ■

Exemples

- On a $(\tan x-\sin x) \underset{x\to 0}{\ll} x^2$ et $(\tan x-\sin x) \underset{x\to 0}{\sim} (1/2)x^3$.
- $\big[\tan x-(\sin x)^2\big] \underset{x\to 0}{\sim} x$.
- Puisque $x^3 \underset{x\to 0}{\ll} x^2$, la propriété (5) montre que $x^2+x^3 \underset{x\to 0}{\sim} x^2$.
  De même, $x^2+(\tan x-\sin x) \underset{x\to 0}{\sim} x^2$, car $\tan x-\sin x$ est infiniment petit devant x^2.
- On sait que $\dfrac{\sin x}{x}$ et $\dfrac{\tan x}{x}$ tendent vers 1 quand x tend vers 0, donc $\sin x \underset{x\to 0}{\sim} x$ et $\tan x \underset{x\to 0}{\sim} x$. Cependant, la différence $\tan x-\sin x$, infiniment petite devant x, n'est pas équivalente à x.

► On a $x \underset{x\to+\infty}{\ll} x^2$ et $\sqrt{x} \underset{x\to+\infty}{\ll} x$.

En effet, $x/x^2 = 1/x$ et $\sqrt{x}/x = 1/\sqrt{x}$ tendent vers 0 quand x tend vers $+\infty$.

Proposition. *Soit une fonction polynôme $P(x) = a_p x^p + a_{p+1}x^{p+1} + \cdots + a_n x^n$, avec $p \leqslant n$, $a_p \neq 0$ et $a_n \neq 0$. Alors on a $P(x) \underset{x\to 0}{\sim} a_p x^p$ et $P(x) \underset{x\to\pm\infty}{\sim} a_n x^n$.*

2.3 Fonctions usuelles

Les fonctions puissances

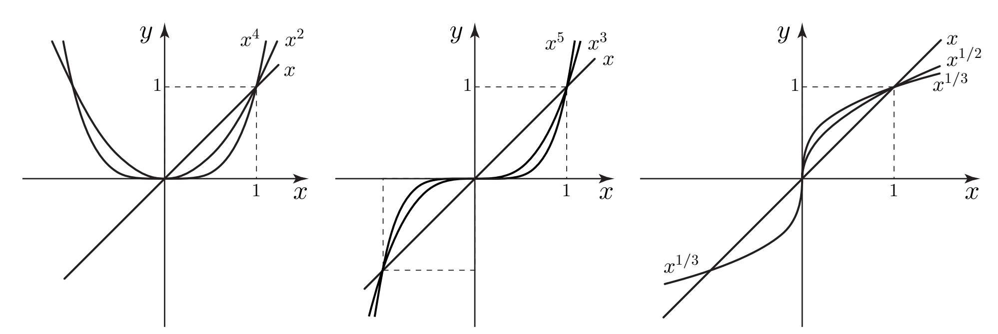
Soit n un entier strictement positif.

► La fonction puissance $x \mapsto x^n$ est définie sur $\mathbb{R}$ et l'on a $\lim\limits_{x\to 0} x^n = 0$, $\lim\limits_{x\to+\infty} x^n = +\infty$.

► Rappelons que si $x \neq 0$, on pose $x^{-n} = 1/x^n$: la fonction $x \mapsto x^{-n}$ est définie sur $\mathbb{R} \setminus \{0\}$. On a $\lim\limits_{x\to 0^+} (x^{-n}) = +\infty$ et $\lim\limits_{x\to+\infty} (x^{-n}) = 0$.

Les fonctions racines

La fonction racine n-ième, notée $x \mapsto \sqrt[n]{x}$ ou $x \mapsto x^{1/n}$, est définie sur $[0, +\infty[$. Par définition, on a les relations $(\sqrt[n]{x})^n = x = \sqrt[n]{x^n}$, ce qui veut dire que les fonctions $x \mapsto \sqrt[n]{x}$ et $x \mapsto x^n$ sont des bijections réciproques sur $[0, +\infty[$.
Les limites aux bornes de l'intervalle sont $\lim\limits_{x\to 0^+} \sqrt[n]{x} = 0$ et $\lim\limits_{x\to+\infty} \sqrt[n]{x} = +\infty$.



La fonction logarithme

La fonction logarithme, notée $x \mapsto \ln x$, est définie sur $]0, +\infty[$. Elle est croissante et non majorée, donc $\lim\limits_{x\to+\infty} \ln x = +\infty$. Quand x tend vers 0 par valeurs positives, $1/x$ tend vers l'infini, donc $\ln x = -\ln(1/x)$ tend vers $-\infty$.

Propriétés du logarithme

a) $\ln 1 = 0$, $(\ln x > 0 \iff x > 1)$, $(\ln x < 0 \iff 0 < x < 1)$, $\ln e = 1$.
b) $\ln(xy) = \ln x + \ln y$, $\ln(x^n) = n \ln x$ et $\ln(x^{1/n}) = (1/n)\ln x$.
c) $\ln'(x) = 1/x$.
d) $\lim\limits_{x\to+\infty} \dfrac{\ln x}{x} = 0$ et $\lim\limits_{x\to 0^+} (x \ln x) = 0$.

La dernière propriété reste vraie si on élève $\ln x$ ou x à des puissances positives :

si p *et* r *sont des exposants positifs, alors* $(\ln x)^p \underset{x\to+\infty}{\ll} x^r$ *et* $x^r(\ln x)^p \underset{x\to 0^+}{\ll} 1$

En effet, $u(x) = \dfrac{\ln x}{x^{r/p}} = \dfrac{p}{r}\dfrac{\ln(x^{r/p})}{x^{r/p}} = \dfrac{p}{r}\dfrac{\ln y}{y}$, où $y = x^{r/p}$ tend vers l'infini avec x ; puisque $\dfrac{\ln y}{y}$ tend vers 0, $u(x)$ aussi. Quand x tend vers 0 par valeurs positives, $z = 1/x$ tend vers $+\infty$, donc $(\ln z)^p/z^r$ tend vers 0 d'après ce qu'on vient de montrer ; comme on a $\ln z = |\ln x|$, il vient $(\ln z)^p/z^r = x^r|\ln x|^p$, donc $\lim\limits_{x\to 0}\left[x^r(\ln x)^p\right] = 0$.

La fonction exponentielle

La fonction exponentielle, notée $x \mapsto \exp x$ ou $x \mapsto e^x$, est définie sur $\mathbb{R}$. C'est la bijection réciproque du logarithme, donc on a $\ln(\exp x) = x$ pour tout x et $\exp(\ln x) = x$ pour tout $x > 0$. La fonction exponentielle est croissante et non majorée, donc $\lim\limits_{x\to+\infty} e^x = +\infty$. Puisque $e^{-x} = 1/e^x$, on en déduit $\lim\limits_{x\to-\infty} e^x = 0$.

Propriétés de l'exponentielle. On les montre en prenant le logarithme.

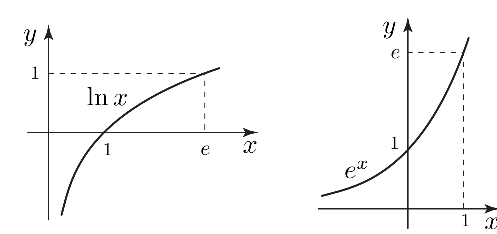
a) $e^0 = 1$, $(e^x > 1 \iff x > 0)$, $(0 < e^x < 1 \iff x < 0)$.

b) $e^{x+y} = e^x e^y$, $e^{nx} = (e^x)^n$ et $e^{x/n} = (e^x)^{1/n}$.

c) $\exp'(x) = \exp x$.

d) Pour tout $r > 0$, on a $\lim\limits_{x\to+\infty}(e^x/x^r) = +\infty$ et $\lim\limits_{x\to-\infty}\left(|x|^r e^x\right) = 0$.

On en déduit que pour tout entier n, on a $x^n \underset{x\to+\infty}{\ll} e^x$ et $e^{-x} \underset{x\to+\infty}{\ll} 1/x^n$.



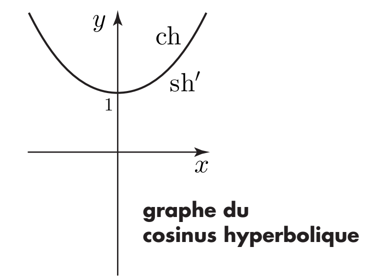
Fonctions hyperboliques

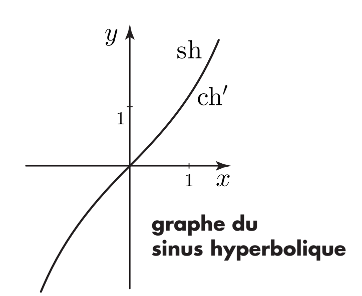
On pose : $\operatorname{ch} x = \dfrac{e^x + e^{-x}}{2}$: c'est la fonction *cosinus hyperbolique,*

$\operatorname{sh} x = \dfrac{e^x - e^{-x}}{2}$: c'est la fonction *sinus hyperbolique.*

► La fonction ch est paire et la fonction sh est impaire.

► On a $(\operatorname{ch} x)^2 - (\operatorname{sh} x)^2 = 1$, $\operatorname{ch}' x = \operatorname{sh} x$ et $\operatorname{sh}' x = \operatorname{ch} x$ pour tout x.





Fonctions puissances généralisées

Si n est un entier, on a $x^n = \exp(n \ln x)$ pour tout $x > 0$. L'expression $\exp(n \ln x)$ a encore un sens si l'on remplace l'entier n par n'importe quel nombre réel α : cela permet de formuler la définition suivante de la puissance α-ième d'un nombre $x > 0$.

Définition

Soit α un nombre réel. Pour tout $x > 0$, on pose $x^\alpha = \exp(\alpha \ln x)$. La fonction $x \mapsto x^\alpha$ s'appelle une *fonction puissance*.

Comme avec des exposants entiers, on a les règles de calcul :

$$\text{pour } x > 0 \text{ et } y > 0, \quad (xy)^\alpha = x^\alpha y^\alpha \ , \ x^{\alpha+\beta} = x^\alpha x^\beta \ , \ (x^\alpha)^\beta = x^{\alpha\beta} \text{ et } x^0 = 1.$$

Voici une propriété utile dont la vérification est immédiate.

Si $u(x)$ et $v(x)$ sont positifs et si $u(x) \underset{x\to a}{\sim} v(x)$, alors $\big[u(x)\big]^\alpha \underset{x\to a}{\sim} \big[v(x)\big]^\alpha$.

2.4 Échelles de comparaison

Une échelle de comparaison quand x tend vers $+\infty$

On a toujours $x^r/x^s = x^{r-s}$; si $r < s$, l'exposant $r-s$ est négatif, donc x^r/x^s tend vers 0 quand x tend vers l'infini :

si $r < s$, alors $x^r \underset{x\to+\infty}{\ll} x^s$.

Puisqu'on a $x^r \underset{x\to+\infty}{\ll} e^x$ et $\ln x \underset{x\to+\infty}{\ll} x^r$ pour $r > 0$, on en déduit le classement suivant.

$$\text{Quand } x \to +\infty \ : \ln x \ll x^{1/n} \ll \cdots \ll x^{1/2} \ll x \ll x^2 \ll \cdots \ll x^n \ll e^x$$

Toutes ces quantités sont infiniment grandes quand x tend vers $+\infty$, mais chacune est infiniment petite devant la suivante.

En prenant les inverses, on obtient une échelle d'infiniments petits, c'est-à-dire de quantités tendant vers 0 quand x tend vers $+\infty$.

$$\text{Quand } x \to +\infty \ : e^{-x} \ll \frac{1}{x^n} \ll \cdots \ll \frac{1}{x^2} \ll \frac{1}{x} \ll \frac{1}{\sqrt{x}} \ll \cdots \ll \frac{1}{\sqrt[n]{x}} \ll \frac{1}{\ln x} \ll 1$$



Une échelle de comparaison quand x tend vers 0

Si $r > s$, $x^r/x^s = x^{r-s}$ tend vers 0 quand x tend vers 0, donc x^r est infiniment petit devant x^s. Nous avons montré que $x^r \ln x$ tend toujours vers 0 quand x tend vers 0, donc x^r est infiniment petit devant $1/\ln x$. On en déduit une échelle d'infiniments petits quand x tend vers 0 par valeurs positives.

$$\text{Quand } x \to 0 \ : \cdots \ll x^3 \ll x^2 \ll x \ll x^{1/2} \ll x^{1/3} \ll \cdots \ll x^{1/n} \ll \frac{1}{\ln x} \ll 1$$

En prenant les inverses, on obtient une échelle d'infiniments grands, c'est-à-dire de quantités tendant vers $+\infty$ quand x tend vers 0 par valeurs positives.

$$\text{Quand } x \to 0 \ : 1 \ll |\ln x| \ll \frac{1}{x^{1/n}} \ll \cdots \ll \frac{1}{\sqrt{x}} \ll \frac{1}{x} \ll \frac{1}{x^2} \ll \frac{1}{x^3} \ll \cdots$$

Exemple. Quel est l'ordre de grandeur de $[x^3-9x+10]^{1/2}$ quand x tend vers 2 ? On a $x^3-9x+10=(x-2)(x^2+2x-5)$ et $\lim_{x\to 2}(x^2+2x-5)=3$, donc $(x^3-9x+10) \underset{x\to 2}{\sim} 3(x-2)$.
On en déduit $\left[x^3-9x+10\right]^{1/2} \underset{\substack{x\to 2\\ x>2}}{\sim} \sqrt{3}\,\sqrt{x-2}$.

Définition
Soit f une fonction. Si l'on a $f(x) \underset{x\to a}{\sim} u(x-a)$, où u une fonction de comparaison en 0, ou un produit de telles fonctions, on dit que $u(x-a)$ est la *partie principale* de $f(x)$ quand x tend vers a. De même, si $f(x) \underset{x\to+\infty}{\sim} v(x)$, alors $v(x)$ est la partie principale de $f(x)$ quand x tend vers $+\infty$.

Croissance exponentielle et factorielle pour les suites

Soit a un nombre réel tel que $a > 1$.

► *a^n tend vers $+\infty$ quand n tend vers $+\infty$ et pour tout nombre α, on a $a^n \underset{n\to+\infty}{\gg} n^\alpha$.*

En effet, puisque $\ln a > 0$, $a^n = e^{n \ln a}$ est infiniment grand devant toute puissance de n, quand n tend vers $+\infty$.

Nous allons montrer $n!$ est infiniment grand devant a^n quand n tend vers $+\infty$ et donner un équivalent utile de $n!$.

► *Pour tout nombre $a > 1$, on a* $n! \underset{n\to+\infty}{\gg} a^n$

► **formule de Stirling :** $n! \underset{n\to+\infty}{\sim} n^n\, e^{-n}\sqrt{2\pi n}$

Choisissons un entier $N > 2a$. Pour tout entier $n = N + p$, on a

$$\frac{a^n}{n!} = \frac{a^N}{N!}\,\frac{a}{N+1}\,\frac{a}{N+2}\cdots\frac{a}{N+p}\,.$$

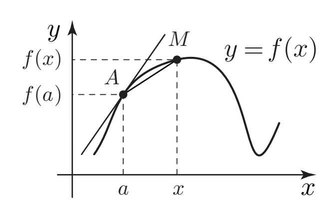
Chacun des rapports $\dfrac{a}{N+i}$ est inférieur à $\dfrac{a}{N} < \dfrac{1}{2}$, donc $\dfrac{a^n}{n!} < \dfrac{a^N}{N!}\,\dfrac{1}{2^p}$. Quand n tend vers l'infini, p aussi, donc $1/2^p$ tend vers 0 et $a^n/n!$ également.

3. La dérivée

3.1 Définition de la dérivée

Soit f une fonction définie sur un intervalle I et soit a un nombre qui est dans I ou bien qui est une extrémité de I.

Définition
Si le rapport $\dfrac{f(x)-f(a)}{x-a}$ a une limite finie quand x tend vers a, cette limite s'appelle la *dérivée* de f en a et se note $f'(a)$.



Sur le graphe de f, plaçons le point A d'abscisse a et un point M d'abscisse $x \neq a$. Le segment de droite AM a pour pente $\dfrac{f(x)-f(a)}{x-a}$. La dérivée de f en a, si elle existe, est par définition la limite des pentes de ces segments AM quand le point M du graphe tend vers A.

Le nombre $f'(a)$ est la pente de la tangente en A au graphe de f.

L'équation de la tangente en A est : $y - f(a) = f'(a)(x-a)$.

La différence $f(x) - f(a)$ est *l'accroissement* des valeurs de la fonction entre a et x : la dérivée en a permet ainsi de comparer, du point de vue de l'ordre de grandeur, l'accroissement de la fonction et l'accroissement de la variable au voisinage de a.

$$\text{Si } f'(a) \neq 0, \text{ alors } \big[f(x)-f(a)\big] \underset{x\to a}{\sim} f'(a)(x-a).$$

Exemple. Calculons la partie principale de $f(x) = \dfrac{(\sin x)^2}{1-e^{-ax}}$ quand x tend vers 0 (a est un nombre différent de 0).
On a $\sin x \underset{x\to 0}{\sim} x$ et comme la dérivée en 0 de $1-e^{-ax}$ est a, il vient $1-e^{-ax} \underset{x\to 0}{\sim} ax$.
On en déduit que $f(x)$ est équivalent à $\dfrac{x^2}{ax} = \dfrac{x}{a}$ quand x tend vers 0.

Proposition. *Pour qu'un nombre k soit la dérivée de f en a, il faut et il suffit que l'on puisse écrire $f(x) = f(a) + k(x-a) + \varphi(x)$, où $\varphi(x) \underset{x\to a}{\ll} (x-a)$.*

Démonstration. On a en effet $\dfrac{f(x)-f(a)}{x-a} - k = \dfrac{\varphi(x)}{x-a}$, donc $\dfrac{f(x)-f(a)}{x-a}$ tend vers k quand x tend vers a si et seulement si $\dfrac{\varphi(x)}{x-a}$ tend vers 0. ■

Approximation affine en un point

Supposons que f est une fonction dérivable en a. Par définition, on a donc

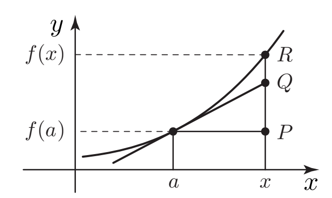
$$f(x) = f(a) + f'(a)(x-a) + \varphi(x), \quad \text{où } \varphi(x) \underset{x\to a}{\ll} (x-a).$$

Pour des valeurs de x assez proches de a, la fonction affine $x \mapsto f(a) + f'(a)(x-a)$ est une approximation de la fonction f à condition de considérer comme négligeable la différence $\varphi(x)$ infiniment petite devant $x - a$.

> **Définition**
> L'application affine $T_a : x \mapsto f(a) + f'(a)(x-a)$ s'appelle *l'approximation affine de* f *au point* a.

Le graphe de T_a a pour équation $y = f(a) + f'(a)(x-a)$:

le graphe de l'approximation affine de f *au point* a *est la tangente en* a *au graphe de* f.



- Sur la figure, l'ordonnée de P est $f(a)$, celle de Q est $T_a(x)$, donc $\overline{PQ} = f'(a)(x-a)$.
- La distance QR est négligeable devant $|x-a|$ pour x assez proche de a.

3.2 Calcul des dérivées

Rappelons les règles pour calculer une dérivée ainsi que les dérivées à connaître.

- $(u+v)' = u' + v'$, $(uv)' = u'v + uv'$, $\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$.
- **dérivée d'une composée** : $(f \circ u)'(x) = f'\big[u(x)\big]u'(x)$.
- $\exp'(x) = \exp x$ et $\ln'(x) = 1/x$.
- La dérivée de la fonction $x \mapsto x^\alpha$ est $\alpha x^{\alpha-1}$ pour tout $x > 0$.
- $\sin'(x) = \cos x$, $\cos'(x) = -\sin x$ et $\tan'(x) = 1 + (\tan x)^2 = \frac{1}{(\cos x)^2}$.

Vérifions que la fonction $f(x) = x^\alpha$ se dérive bien comme une fonction puissance entière. Par définition, on a $f(x) = \exp\big[u(x)\big]$, où $u(x) = \alpha \ln x$. D'après la règle pour dériver une composée, il vient donc

$f'(x) = \exp'\big[u(x)\big]u'(x) = \exp\big[u(x)\big]\frac{\alpha}{x} = \alpha\exp\big[\alpha\ln x\big]\exp\big[-\ln x\big] = \alpha\exp\big[(\alpha-1)\ln x\big] = \alpha x^{\alpha-1}$.

Exemples d'approximations affines

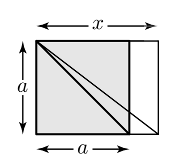
Dans les formules ci-dessous, $\varphi(x)$ désigne des quantités infiniment petites devant x quand x tend vers 0, autrement dit $\varphi(x) \underset{x\to 0}{\ll} x$.

- $(1+x)^\alpha = 1 + \alpha x + \varphi(x)$ pour tout nombre α positif ou négatif,
- $e^x = 1 + x + \varphi(x)$ et $\ln(1+x) = x + \varphi(x)$
- $\sin x = x + \varphi(x)$, $\cos x = 1 + \varphi(x)$ et $\tan x = x + \varphi(x)$.

En effet, les fonctions e^x, $\ln(1+x)$, $\sin x$ et $\tan x$ ont pour dérivée 1 en $x = 0$, la dérivée en 0 de $(1+x)^\alpha$ est α et $\cos'(0) = -\sin 0 = 0$.

- Pour u assez petit, le sinus d'un angle de u degrés est équivalent à $\frac{\pi}{180}u$.

Exemple. Considérons un carré de côté a qu'on déforme en un rectangle en allongeant très légèrement deux côtés opposés : comment varie la diagonale ?

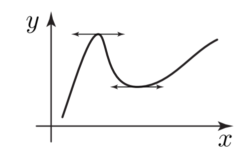


La diagonale du carré est $\sqrt{a^2+a^2} = a\sqrt{2}$ et celle du rectangle est $f(x) = \sqrt{a^2+x^2}$ en appelant x la longueur du grand côté du rectangle. On a $f'(x) = \dfrac{x}{\sqrt{a^2+x^2}}$, donc l'approximation affine de f en a est $f(a)+f'(a)(x-a) = \sqrt{2a^2} + \dfrac{a}{\sqrt{2a^2}}(x-a) = a\sqrt{2} + \dfrac{x-a}{\sqrt{2}}$. Pour un petit allongement $\ell = x-a$ du côté, la diagonale du carré s'allongera d'environ $\ell/\sqrt{2}$.

3.3 Comportement d'une fonction au voisinage d'un point

Une fonction f a un *maximum local* au point a si pour tout x assez proche de a, $f(x)$ est inférieur ou égal à $f(a)$. De même, f a un *minimum local* en a si l'on a $f(x) \geqslant f(a)$ pour tout x assez proche de a.

Proposition. *Si f est dérivable en a et si f a un maximum ou un minimum local en a, alors $f'(a) = 0$.*



La proposition signifie qu'en un point où f a un maximum ou un minimum local, la tangente est horizontale, à condition que f soit dérivable en ce point.

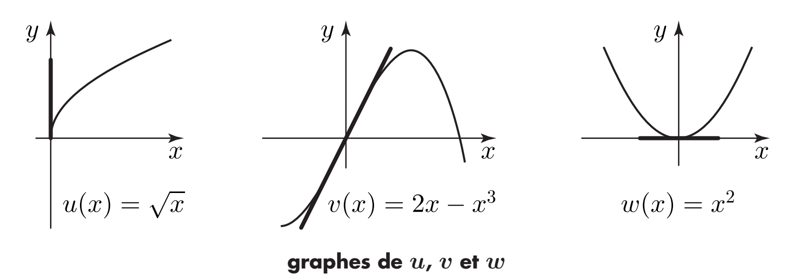
Démonstration. Si $f'(a) \neq 0$, on sait que $f(x) - f(a)$ a le signe de $f'(a)(x-a)$ pour tout x assez proche de a. Puisque $x-a$ change de signe en a, c'est que la différence $f(x) - f(a)$ ne garde pas un signe constant autour de a. Cela montre que si $f'(a) \neq 0$, alors f n'a pas d'extremum local en a. ■

La condition $f'(a) = 0$ n'entraîne pas que f a un maximum ou un minimum local : ainsi la dérivée de $x \mapsto x^3$ est nulle en $x = 0$, mais cette fonction n'a pas d'extremum en 0 comme le montre le graphe page 267.

Formulons encore deux propriétés utiles.

- Si $f'(a) = 0$, alors $\big[f(x) - f(a)\big] \underset{x \to a}{\ll} (x-a)$.
- Si $\displaystyle\lim_{x \to a} \frac{f(x) - f(a)}{x-a} = \pm\infty$, alors $\big|f(x) - f(a)\big| \underset{x \to a}{\gg} |x-a|$.

Exemples. Voici trois fonctions u, v, w qui tendent vers 0 quand x tend vers 0 mais qui se comportent bien différemment au voisinage de 0 :



graphes de u, v et w

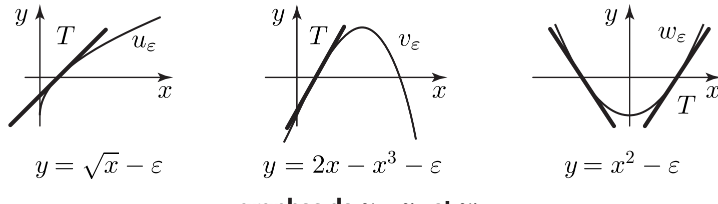
► La fonction $u(x) = \sqrt{x}$ n'est pas dérivable en 0, car le rapport $\sqrt{x}/x = 1/\sqrt{x}$ tend vers $+\infty$ quand x tend vers 0 par valeurs positives : on a $u(x) \underset{x \to 0^+}{\gg} x$, autrement dit, au voisinage de l'origine, l'accroissement de la fonction est infiniment grand devant celui de la variable. Intuitivement, la pente de la tangente à l'origine est infinie, ce qui signifie qu'à l'origine, le graphe est tangent à l'axe des ordonnées.

► Pour la fonction $v(x) = 2x - x^3$, la tangente à l'origine a pour pente $v'(0) = 2$: on a donc $v(x) \underset{x \to 0}{\sim} 2x$.

► En 0, la fonction $w(x) = x^2$ a pour dérivée $w'(0) = 0$: le graphe de w est tangent à l'axe des abscisses. On a $w(x) \underset{x \to 0}{\ll} x$ et quand x tend vers 0, l'accroissement de la fonction est infiniment petit devant celui de la variable.

Un comportement comme celui de u ou de w ne se produit qu'en des points exceptionnels : lorsqu'une courbe dérivable coupe l'axe des abscisses, il y a en général au point d'intersection une tangente de pente finie non nulle.

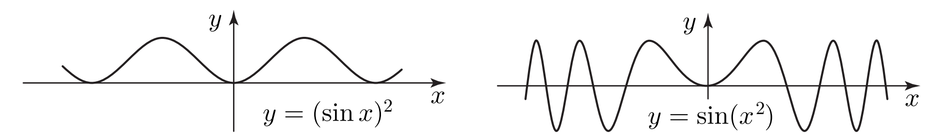
Par exemple, perturbons un peu les fonctions u et w en leur soustrayant un petit nombre $\varepsilon > 0$. Les graphes des fonctions $u_\varepsilon(x) = u(x) - \varepsilon$ et $w_\varepsilon(x) = w(x) - \varepsilon$ s'obtiennent à partir des graphes ci-dessus en remontant de ε l'axe des abscisses.



graphes de u_ε, v_ε et w_ε

Cette fois, le graphe de u_ε coupe l'axe des abscisses en un point à tangente T de pente finie non nulle et il en va de même pour le graphe de w_ε. Cette propriété était vraie pour v et elle reste vraie pour $v_\varepsilon(x) = v(x) - \varepsilon$.

Exemples. Voici les graphes de la fonction $(\sin x)^2$ et de la fonction $\sin(x^2)$:



- Puisque $(\sin x)^2$ a pour dérivée $2\sin x\cos x$, le graphe de $(\sin x)^2$ est tangent à l'axe des x en tout point $k\pi$, où le sinus s'annule. Le maximum est atteint aux points $(\pi/2)+k\pi$, où $\sin x=\pm 1$. La fonction est périodique de période π, car $\sin(x+\pi)=-\sin x$, donc $\big(\sin(x+\pi)\big)^2=(\sin x)^2$.
- Pour le graphe de $\sin(x^2)$, on a $\sin(x^2)\sim x^2$ quand x tend vers 0, d'où la forme parabolique au voisinage de l'origine. Les oscillations sont de plus en plus rapides au fur et à mesure qu'on s'éloigne de l'origine : en effet, $\sin(x^2)$ s'annule aux points $x_k=\pm\sqrt{k\pi}$, où $k\in\mathbb{N}$, et entre deux zéros consécutifs, l'écart
$$|x_{k+1}-x_k|=\sqrt{(k+1)\pi}-\sqrt{k\pi}=\frac{\pi}{\sqrt{(k+1)\pi}+\sqrt{k\pi}}$$
décroît en tendant vers 0.

3.4 Graphes du carré et de la racine carrée d'une fonction

Supposons qu'on connaisse le graphe d'une fonction dérivable f et qu'on veuille dessiner rapidement l'allure des fonctions $\sqrt{f(x)}$ et $\big[f(x)\big]^2$.
Le domaine de définition de $\sqrt{f(x)}$ est formé des nombres x tels que $f(x)\geqslant 0$: ils correspondent aux points du graphe de f situés au dessus de l'axe des abscisses.

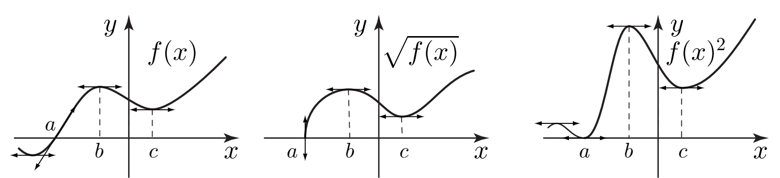
Règles à suivre pour dessiner ces graphes

1) *Les fonctions $\sqrt{f(x)}$ et $\big[f(x)\big]^2$ s'annulent exactement aux mêmes points que f.*
2) *Sur tout intervalle où f est positif ou nul, les fonctions $\sqrt{f(x)}$ et $\big[f(x)\big]^2$ varient dans le même sens que f.*

   Si par exemple on a $0\leqslant f(x)<f(y)$, alors $\sqrt{f(x)}<\sqrt{f(y)}$ et $\big[f(x)\big]^2<\big[f(y)\big]^2$.
3) *Sur un intervalle où $f(x)\leqslant 0$, les variations de $\big[f(x)\big]^2$ sont opposées à celles de f.*
4) *Si le graphe de f coupe l'axe des abscisses avec une tangente de pente non nulle, alors en ce point le graphe de $\sqrt{f(x)}$ a une tangente verticale.*
5) *Si le graphe de f coupe l'axe des abscisses avec une tangente non verticale, alors en ce point, le graphe de $\big[f(x)\big]^2$ est tangent à l'axe des abscisses.*

   En notant a ce point, on a $f(a)=0$ et si par exemple $f'(a)>0$, $f(x)$ est positif pour $x>a$ assez voisin de a : $\sqrt{f(x)}$ est donc défini sur un petit intervalle $J=\,]a,v[$. Pour $x\in J$, on a $\dfrac{\sqrt{f(x)}}{x-a}=\dfrac{1}{\sqrt{f(x)}}\dfrac{f(x)}{x-a}$. On en déduit que si le rapport $\dfrac{f(x)}{x-a}$ a une limite non nulle quand x tend vers a, alors $\dfrac{\sqrt{f(x)}}{x-a}$ tend vers l'infini, car

$\lim\limits_{x\to a} 1/\sqrt{f(x)} = +\infty$: cela veut dire que le graphe de $\sqrt{f(x)}$ a une tangente verticale en a. La dérivée de $\left[f(x)\right]^2$ au point a est $2f(a)f'(a) = 0$ puisque $f(a) = 0$: le graphe de $\left[f(x)\right]^2$ est donc tangent en a à l'axe des abscisses.



6) *Si* $f(x) \underset{x\to+\infty}{\ll} x^2$*, alors quand* x *tend vers* $+\infty$*, le graphe de* $\sqrt{f(x)}$ *a pour direction asymptotique l'axe des abscisses.*

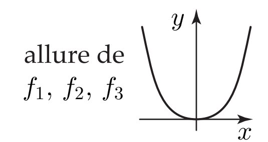
Il suffit de remarquer que si $f(x)/x^2$ tend vers 0 quand x tend vers $+\infty$, alors $\sqrt{f(x)}/x = \sqrt{f(x)/x^2}$ tend aussi vers 0.

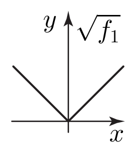
7) *Si* $f(x) \underset{x\to+\infty}{\gg} \sqrt{x}$*, alors quand* x *tend vers* $+\infty$*, le graphe de* $\left[f(x)\right]^2$ *a pour direction asymptotique l'axe des ordonnées.*

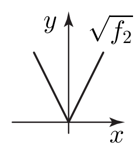
Remarque

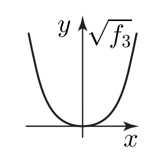
Si le graphe de f est tangent à l'axe des abscisses en un point a, on ne peut rien en déduire sur le comportement de $\sqrt{f(x)}$ au voisinage de a. Ainsi les fonctions $f_1(x)=x^2$, $f_2(x) = 4x^2$ et $f_3(x) = 4x^4$ sont tangentes à l'axe des x à l'origine, mais en ce point

- $\sqrt{f_1(x)} = |x|$ a deux demi-tangentes de pente 1 et -1 ;
- $\sqrt{f_2(x)} = 2|x|$ a deux-demi tangentes de pente 2 et -2 ;
- $\sqrt{f_3(x)} = 2x^2$ est tangente à l'axe des x.





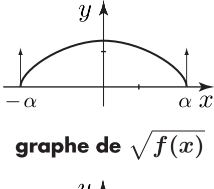




Un exemple. Dessinons la courbe d'équation $y^2 = 0{,}7 + \cos x$, où $-\pi \leqslant x \leqslant \pi$.
Posons $f(x)=0{,}7+\cos x$. Puisque l'équation de la courbe est $y=\pm\sqrt{f(x)}$, commençons par dessiner le graphe de la fonction $\sqrt{f(x)} = \sqrt{0{,}7 + \cos x}$ entre $-\pi$ et π.
Il y a un nombre $\alpha \in]0,\pi[$ tel que $\cos\alpha = -0{,}7$ (α est peu différent de 2,34). On a $f(\pm\alpha) = 0$. Puisqu'on a $\cos x > \cos\alpha$ si $-\alpha < x < \alpha$, f prend des valeurs positives ou nulles sur $[-\alpha,\alpha]$: c'est l'intervalle de définition de $\sqrt{f(x)}$.
Sur $[0,\alpha]$, f décroît de 1,7 à 0 ; comme f est paire, on en déduit qu'elle est croissante sur $[-\alpha, 0]$.

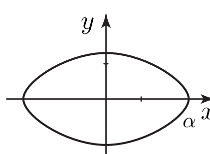
graphe de $f(x)$

- Quand x va de $-\alpha$ à 0, $\sqrt{f(x)}$ croît de 0 à $\sqrt{1{,}7}$; quand x va de 0 à α, $\sqrt{f(x)}$ décroît de $\sqrt{1{,}7}$ à 0.
- On a $f'(x) = -\sin x$ donc au point α, le graphe de f a une tangente de pente $-\sin\alpha \neq 0$. Par conséquent, $\sqrt{f(x)}$ a une tangente verticale en α, et de même en $-\alpha$.



graphe de $\sqrt{f(x)}$

Pour dessiner la courbe d'équation $y^2 = 0{,}7 + \cos x$ entre $-\pi$ et π, il suffit de compléter le graphe de $\sqrt{f(x)}$ par symétrie par rapport à l'axe des abscisses. On trouve ainsi une courbe fermée et symétrique par rapport à chacun des axes.



3.5 Les fonctions trigonométriques réciproques

La fonction Arc sinus

La fonction $x \mapsto \sin x$ définit une bijection croissante $[-\pi/2, \pi/2] \to [-1, 1]$: pour tout nombre $y \in [-1, 1]$, il existe un unique nombre $x \in [-\pi/2, \pi/2]$ tel que $y = \sin x$.

Définition

Pour tout $y \in [-1, 1]$, on note $\operatorname{Arc\,sin} y$ le nombre $x \in [-\pi/2, \pi/2]$ tel que $y = \sin x$; ce nombre s'appelle *l'Arc sinus de y*. La fonction $y \mapsto \operatorname{Arc\,sin} y$ est une bijection croissante et impaire de $[-1, 1]$ sur $[-\pi/2, \pi/2]$.

On a $\operatorname{Arc\,sin} 0 = 0$, $\operatorname{Arc\,sin}\frac{1}{2} = \frac{\pi}{6}$, $\operatorname{Arc\,sin}\frac{\sqrt{2}}{2} = \frac{\pi}{4}$, $\operatorname{Arc\,sin}\frac{\sqrt{3}}{2} = \frac{\pi}{3}$ et $\operatorname{Arc\,sin} 1 = \frac{\pi}{2}$.

Pour dériver $f(y) = \operatorname{Arc\,sin} y$, écrivons que pour tout $y \in \,]-1, 1[$, on a l'identité $\sin\big(f(y)\big) = y$ et dérivons en appliquant la règle pour une fonction composée. Il vient $\big[\cos\big(f(y)\big)\big] f'(y) = 1$. Puisqu'on a supposé y strictement compris entre -1 et 1, le nombre $x = f(y)$ est dans l'intervalle ouvert $]-\pi/2, \pi/2[$, donc $\cos x > 0$. Ainsi on a $\cos x = \sqrt{1 - (\sin x)^2} = \sqrt{1 - y^2}$, car $\sin x = y$, et il vient

$$\operatorname{Arc\,sin}'(y) = \frac{1}{\sqrt{1 - y^2}}, \quad \text{pour tout } y \in \,]-1, 1[.$$

La fonction Arc tangente

La fonction $x \mapsto \tan x$ définit une bijection croissante $]-\pi/2, \pi/2[\to]-\infty, +\infty[$: pour tout nombre réel y, il existe un unique nombre $x \in \,]-\pi/2, \pi/2[$ tel que $y = \tan x$.

Définition

Pour tout $y \in \mathbb{R}$, on note $\operatorname{Arc\,tan} y$ le nombre $x \in \,]-\pi/2, \pi/2[$ tel que $y = \tan x$; ce nombre s'appelle *l'Arc tangente de y*. La fonction $y \mapsto \operatorname{Arc\,tan} y$ est une bijection croissante et impaire de $\mathbb{R}$ sur $]-\pi/2, \pi/2[$.

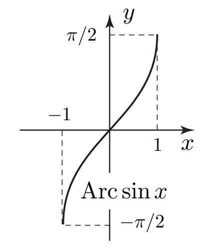
On a $\operatorname{Arc\,tan} 0 = 0$, $\operatorname{Arc\,tan} 1 = \pi/4$ et $\lim_{x \to +\infty} (\operatorname{Arc\,tan} x) = \pi/2$.

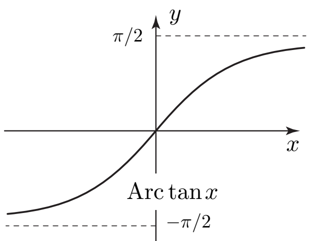
Pour dériver $g(y) = \operatorname{Arc}\tan y$, écrivons que pour tout y, on a $\tan\big(g(y)\big) = y$: on obtient $\big[\tan'\big(g(y)\big)\big]g'(y) = 1$. Posons $x = g(y)$, de sorte que $\tan x = y$. On a

$$\tan'(x) = 1 + (\tan x)^2 = 1 + y^2 ,$$

d'où $(1 + y^2)g'(y) = 1$.

$$\operatorname{Arc}\tan'(y) = \frac{1}{1+y^2} , \quad \text{pour tout } y \in \mathbb{R}.$$





3.6 La différentielle

Supposons que f est une fonction dérivable en a. Par définition, l'accroissement $f(x) - f(a)$ est alors approché par $f'(a)(x - a)$, avec une erreur infiniment petite devant $x-a$.

> **Définition**
> La *différentielle en* a de la fonction $y = f(x)$ est la fonction linéaire $t \mapsto f'(a)\,t$. On note dx la variable et dy la valeur de la différentielle, de sorte qu'on a $dy = f'(a)dx$.
> La dérivée de f au point a se note aussi $\dfrac{dy}{dx} = f'(a)$.

Si l'on donne à la variable dx une valeur numérique assez petite δx, alors $\delta y = f'(a)\delta x$ est une bonne approximation de l'accroissement $f(a + \delta x) - f(a)$ des valeurs de la fonction. Traduisons le calcul des dérivées en un calcul des différentielles.
Soient $y = f(x)$ et $z = g(x)$ des fonctions dérivables d'une même variable x.

Somme : $d(y + z) = dy + dz$.

Produit par une constante : Si k est une constante, alors $d(ky) = k\,dy$.

Produit : $d(yz) = (dy)z + y(dz)$.

Quotient : Aux points où $z \neq 0$, $d\left[\dfrac{y}{z}\right] = \dfrac{dy}{z} - \dfrac{y}{z^2}dz$.

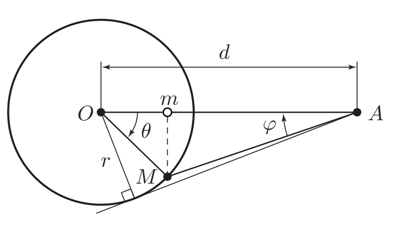
Composée : Si $v = g(y)$ et $y = f(x)$, alors $dv = g'(y)dy = g'(y)f'(x)dx$, ou encore

$$\frac{dv}{dx} = \frac{dv}{dy}\frac{dy}{dx} \qquad \textit{(règle de différentiation en chaîne)}$$

C'est la formule pour dériver la fonction composée $g \circ f$: $\dfrac{dv}{dx} = (g \circ f)'(x) = g'(y)f'(x)$.

Fonctions réciproques l'une de l'autre : Si des quantités x et y sont fonctions réciproques l'une de l'autre, leurs dérivées en des points qui se correspondent sont inverses l'une de l'autre : $\dfrac{dx}{dy} = \left[\dfrac{dy}{dx}\right]^{-1}$.

Exemple. Un point mobile M décrit un cercle de centre O et de rayon r. On le filme depuis un point fixe A situé à une distance $d > r$ de O. Quelle relation différentielle y a-t-il entre l'angle $\theta = \widehat{\overrightarrow{OA}, \overrightarrow{OM}}$ qui repère la position de M et l'angle $\varphi = \widehat{\overrightarrow{AO}, \overrightarrow{AM}}$ que fait la caméra avec $\overrightarrow{AO}$?



Notons m la projection de M sur OA. On a $Om = r\cos\theta$, et $mM = r\sin\theta$ donc $\tan\varphi = \dfrac{mM}{Am} = \dfrac{r\sin\theta}{d - r\cos\theta}$. L'angle φ est maximum quand AM est tangent au cercle, pour une valeur $\varphi_{\max}$ telle que $\sin(\varphi_{\max}) = \dfrac{r}{d}$. En posant $a = r/d$, il vient

$$\tan\varphi = \frac{a\sin\theta}{1 - a\cos\theta}$$

Les angles φ et θ sont des fonctions du temps qu'on suppose dérivables.

- La différentielle de $\tan\varphi$ est $(1 + \tan^2\varphi)\, d\varphi$,
- et la différentielle de $\dfrac{\sin\theta}{1 - a\cos\theta}$ est $\dfrac{\cos\theta\, d\theta(1 - a\cos\theta) - \sin\theta(a\sin\theta\, d\theta)}{(1 - a\cos\theta)^2}$, donc

(1) $$(1 + \tan^2\varphi)\, d\varphi = a\frac{\cos\theta - a}{(1 - a\cos\theta)^2}\, d\theta$$

On en déduit que la dérivée de la fonction $\theta \mapsto \varphi(\theta)$ est $\dfrac{d\varphi}{d\theta} = \dfrac{a}{1 + \tan^2\varphi}\,\dfrac{\cos\theta - a}{(1 - a\cos\theta)^2}$.

L'angle θ est une fonction $\theta(t)$ du temps. En divisant par dt dans (1) et en notant comme d'habitude $\dot\varphi = \dfrac{d\varphi}{dt}$ et $\dot\theta = \dfrac{d\theta}{dt}$ les dérivées par rapport au temps, on obtient

$$(1 + \tan^2\varphi)\dot\varphi(t) = a\frac{\cos\theta - a}{(1 - a\cos\theta)^2}\dot\theta(t)$$

Par exemple, pour $\theta = 0$, l'angle φ est nul et à cet instant, on a l'égalité $\dot\varphi = \frac{a}{1-a}\dot\theta$. Quand $\varphi = \varphi_{\max}$, on a $\theta = \frac{\pi}{2} - \varphi_{\max}$, donc $\cos\theta = \sin(\varphi_{\max}) = a$ et $\dot\varphi = 0$.

Dérivée logarithmique

Définition

Soit f une fonction dérivable à valeurs strictement positives. La *dérivée logarithmique de* f est la dérivée de la fonction $x \mapsto \ln\big(f(x)\big)$, c'est-à-dire $\dfrac{f'(x)}{f(x)}$.

En posant $z = f(x)$, la différentielle de $\ln z$ est $\dfrac{dz}{z}$ et l'on a les règles de calcul suivantes.

Produit par une constante : Si k est une constante, alors $\dfrac{d(ky)}{ky} = \dfrac{dy}{y}$.

Produit de fonctions : Si $p = yz$, alors $\dfrac{dp}{p} = \dfrac{dy}{y} + \dfrac{dz}{z}$.

Quotient de fonctions : Si $q = \frac{y}{z}$, alors $\frac{dq}{q} = \frac{dy}{y} - \frac{dz}{z}$.

Puissance d'une fonction : Si $u = y^\alpha$, alors $\frac{du}{u} = \alpha \frac{dy}{y}$.

Ces formules s'obtiennent en dérivant dans les relations $\ln(ky) = \ln k + \ln y$, $\ln(yz) = \ln y + \ln z$, $\ln \frac{y}{z} = \ln y - \ln z$ et $\ln(y^\alpha) = \alpha \ln y$. De même, si $p = yz$, alors $dp = (dy)z + y(dz)$, donc $\frac{dp}{p} = \frac{(dy)z + y(dz)}{yz} = \frac{dy}{y} + \frac{dz}{z}$.

Exemple. On mesure le volume d'une sphère avec une erreur relative d'au plus 10%. Comment estimer l'erreur que l'on commet en calculant le rayon r_0 ?

Le volume d'une sphère de rayon r est $v = (4/3)\pi r^3$. On a $\frac{dv}{v} = \frac{d(r^3)}{r^3} = 3\frac{dr}{r}$, donc $\frac{dr}{r} = \frac{1}{3}\frac{dv}{v}$. Si $\frac{dv}{v}$ vaut au plus $1/10$, l'erreur relative sur le rayon est de l'ordre de $1/30$. Il vient donc $\frac{|r - r_0|}{r_0} \leqslant \frac{1}{30}$ ou encore $\frac{29}{30} r_0 \leqslant r \leqslant \frac{31}{30} r_0$.

Coefficient d'élasticité. Supposons que des quantités x et y varient en fonction d'un paramètre en ayant des ordres de grandeur très différents : dans ce cas, la mesure de la dérivée $\frac{dy}{dx}$ n'a pas grand sens.

Mais si l'on considère des variations relatives $\frac{\delta y}{y}$ et $\frac{\delta x}{x}$, ces rapports ne sont pas affectés par un changement d'échelle du type $x \mapsto Kx$ ou $y \mapsto Ky$. Il en va de même des dérivées logarithmiques $\frac{dy}{y}$ et $\frac{dx}{x}$, appelées *élasticités* de y et de x. Pour étudier l'effet produit sur y par une variation de x, on définit le *coefficient d'élasticité* de y par rapport à x en posant :

$$e_{y/x} = \frac{dy/y}{dx/x} = \left(\frac{dy}{dx}\right)\left(\frac{x}{y}\right)$$

Par exemple, dans un marché de consommation, la quantité y demandée pour un bien donné dépend du prix p de ce bien. Certains modèles économiques proposent entre p et y une relation de la forme $y = Ap^n$, où A est une constante positive et n un entier positif ou négatif. On a alors $\frac{dy}{y} = n\frac{dp}{p}$ et $e_{y/p} = \frac{dy/y}{dp/p} = n$. Le coefficient d'élasticité est constant et vaut toujours n : on dit que la relation entre p et y est isoélastique.

Dans la plupart des cas, ce coefficient d'élasticité $e_{y/p}$ est négatif : en effet, une diminution du prix augmente en général la demande, de sorte que la fonction $p \mapsto y$ est décroissante ; pour une fonction puissance $y = p^n$, cela veut dire que l'exposant n est négatif (on dit que le bien est typique). Cependant, certains biens de luxe (qualifiés d'atypiques) ont un coefficient d'élasticité positif.

4. Fonctions continues

Soit f une fonction à valeurs réelles définie sur un intervalle I et soit $a \in I$. Pour être certain qu'en prenant des valeurs de plus en plus proches de a, on obtient des valeurs qui tendent vers $f(a)$, il n'y a pas besoin que la fonction soit dérivable : il suffit que $f(x)$ ait pour limite $f(a)$ quand x tend vers a.

Définition

La fonction f est *continue en* a si $\lim_{x\to a} f(x) = f(a)$. Si f est continue en tout point de I, on dit que f est continue sur I.

La fonction f est continue en a si pour tout nombre $\varepsilon > 0$, il existe un intervalle $[x_0 - \alpha, x_0 + \alpha]$, avec $\alpha > 0$, sur lequel f ne prend que des valeurs comprises entre $f(a) - \varepsilon$ et $f(a) + \varepsilon$.

La propriété de continuité ne permet cependant pas de comparer les ordres de grandeur des écarts $|f(x) - f(a)|$ et $|x - a|$ pour x proche de a.

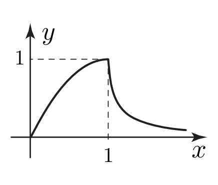
Exemple. Définissons une fonction f en posant

$$f(x) = 2x - x^2 \text{ si } 0 \leqslant x \leqslant 1 \quad \text{et} \quad f(x) = \frac{1}{10x - 9} \text{ si } x > 1.$$

Cette fonction est continue en 1 : en effet, on a $\lim_{x\to 1^-} f(x) = \lim_{x\to 1^-} (2x - x^2) = 1$ et $\lim_{x\to 1^+} f(x) = \lim_{x\to 1^+} \frac{1}{10x-9} = 1$, donc la limite à gauche est égale à la limite à droite, la valeur commune étant $f(1) = 1$.

- Pour $h<0$, on a $f(1+h) = 2(1+h) - (1+h)^2 = 1 - h^2$, donc le rapport $\frac{f(1+h) - f(1)}{h} = \frac{-h^2}{h}$ tend vers 0 quand h tend vers 0 par valeurs négatives : la dérivée à gauche en 1 vaut 0, de sorte que l'accroissement $f(1+h) - f(1)$ est négligeable devant h quand h tend vers 0 par valeurs négatives.
- Pour $h>0$, on a $f(1+h) = \frac{1}{10(1+h)-9} = [1+10h]^{-1} = 1 - 10h + \varphi(h)$, où $\varphi(h) \underset{h\to 0^+}{\ll} h$: la dérivée à droite en 1 vaut donc -10 et quand h tend vers 0 par valeurs positives, on a $f(1+h) - f(1) \sim -10h$.

Puisque la dérivée à gauche en 1 n'est pas égale à la dérivée à droite, la fonction f n'est pas dérivable en 1. Le graphe possède deux demi-tangente au point $(1, 1)$, l'une de pente 0, l'autre de pente -10. Selon le signe de h, les quantités $\big[f(1+h) - f(1)\big]/h$ ont des ordres de grandeur très différents.



Propriétés des fonctions continues

a) *Si des fonctions* f *et* g *sont continues en* a, *leur somme* $x \mapsto f(x)+g(x)$ *et leur produit* $x \mapsto f(x)g(x)$ *sont continus en* a. *Si de plus* $g(a) \neq 0$, *le quotient* $x \mapsto \dfrac{f(x)}{g(x)}$ *est continue en* a.

b) *Supposons que* f *est une fonction ayant pour limite* a *quand* x *tend vers* x_0. *Si* g *est continue en* a, *alors* $g\big(f(x)\big)$ *tend vers* $g(a)$ *quand* x *tend vers* x_0.

c) *Si une suite* (u_n) *a pour limite* a *et si* g *est continue en* a, *alors* $g(u_n)$ *tend vers* $g(a)$.

> Supposons que f est une fonction continue en x_0 et que g est une fonction continue en $a = f(x_0)$. Alors $f(x)$ tend vers $a = f(x_0)$ quand x tend vers x_0, donc (propriété (b)), $g\big(f(x)\big)$ tend vers $g(a) = g\big(f(x_0)\big)$ quand x tend vers x_0. Cela veut dire que la composée $x \mapsto (g \circ f)(x)$ est continue en x_0.

En particulier, si f et g sont continues en tout point et si la composée $g \circ f$ est définie, alors $g \circ f$ est continue en tout point.

La composée de deux fonctions continues est continue.

Proposition. *Une fonction dérivable en* a *est continue en* a.

Démonstration. Si f est une fonction dérivable en a, alors $f(x)-f(a)=f'(a)(x-a)+\varphi(x)$, où la fonction $\varepsilon(x) = \dfrac{\varphi(x)}{x-a}$ tend vers 0 quand x tend vers a. Le produit $(x-a)\varepsilon(x) = \varphi(x)$ tend vers 0 quand x tend vers a, de même que $f'(a)(x-a)$, donc aussi $f(x)-f(a)$: cela veut dire que l'on a $\lim\limits_{x \to a} f(x) = f(a)$. ■

Fonctions continues sur un intervalle

Les fonctions qui sont continues sur tout un intervalle ont des propriétés fondamentales qu'on utilise constamment. Rappelons-les sans démonstration.

Théorème des valeurs intermédiaires. *Soit* f *une fonction à valeurs réelles continue sur l'intervalle* I. *Si* a *et* b *sont des éléments de* I *et si* k *est un nombre compris entre* $f(a)$ *et* $f(b)$, *il existe un nombre* c *compris entre* a *et* b *tel que* $f(c) = k$.

Application. *Soit* f *une fonction à valeurs réelles continue sur un intervalle* I. *Si* a *et* b *sont des éléments de* I *tels que* $f(a)$ *et* $f(b)$ *sont de signes contraires, il existe un nombre* $c \in]a,b[$ *tel que* $f(c) = 0$.

C'est un énoncé utile pour montrer qu'une équation $f(x) = 0$ possède au moins une solution.

> Pour le démontrer, on applique simplement le théorème des valeurs intermédiaires en choisissant la valeur $k = 0$ qui, par hypothèse, est bien comprise entre $f(a)$ et $f(b)$.

Exemple. Voici un tableau de quelques valeurs pour les fonctions $x \mapsto x^2$ et $x \mapsto (3/2)^x$ (rappelons qu'on a posé $a^x = \exp(x \ln a)$ si a est un nombre strictement positif). Puisqu'on a $(3/2)^x = \frac{3^x}{2^x}$, il n'est pas difficile de calculer ces valeurs lorsque x est un entier.

x	0	1	2	12	13
x^2	0	1	4	144	169
$(3/2)^x$	1	1,5	2,25	129,7	194,6

Les fonctions x^2 et $(3/2)^x$ sont toutes deux continues sur l'intervalle $[0, +\infty[$.
Pour $x = 1$, on a $x^2 < (3/2)^x$ et pour $x = 2$, on a $x^2 > (3/2)^x$. La différence $x^2 - (3/2)^x$ change de signe entre 1 et 2 et c'est une fonction continue sur $[0, +\infty[$. D'après l'application du théorème des valeurs intermédiaires, il y a donc un nombre $\alpha \in]1, 2[$ tel que $\alpha^2 = (3/2)^\alpha$.
Pour $x = 12$, on a $x^2 > (3/2)^x$ et pour $x = 13$, on a $x^2 < (3/2)^x$. A nouveau, la différence $x^2 - (3/2)^x$ change de signe entre 12 et 13, donc il y a un nombre $\beta \in]12, 13[$ tel que $\beta^2 = (3/2)^\beta$.
Ainsi l'équation $x^2 = (3/2)^x$ possède au moins les deux solutions α et β ; en étudiant plus précisément ces fonctions, on peut montrer qu'il n'y a pas d'autres solutions.

Corollaire. *Tout polynôme à coefficients réels et de degré impair possède au moins une racine réelle.*

Démonstration. Soit P un polynôme à coefficients réels et de degré n impair. Si le coefficient dominant est par exemple positif, alors $\lim_{x \to +\infty} P(x) = +\infty$ et $\lim_{x \to -\infty} P(x) = -\infty$. Il s'ensuit que $P(x)$ est positif pour des valeurs de x positives assez grandes et que $P(x)$ est négatif pour des valeurs négatives de x assez grandes en valeur absolue. Puisque la fonction $x \mapsto P(x)$ est continue sur $\mathbb{R}$, l'équation $P(x) = 0$ possède au moins une solution dans $\mathbb{R}$. ■

Fonction continue monotone. *Une fonction continue et strictement monotone sur un intervalle I définit une bijection de I sur l'intervalle image $f(I)$ et la bijection réciproque est continue.*

Fonction continue sur un segment. *Soit f une fonction continue sur un segment $[a, b]$.*

a) *Quand x parcourt $[a, b]$, les valeurs $f(x)$ ont un maximum et un minimum : il existe des éléments u et v de $[a, b]$ tels que, pour tout $x \in [a, b]$, on a $f(u) \leqslant f(x) \leqslant f(v)$. Le nombre $m = f(u)$ est le minimum de f, le nombre $M = f(v)$ est le maximum.*

b) *Étant donné un nombre $\varepsilon > 0$, on peut partager le segment $[a, b]$ en sous-intervalles de longueurs assez petites pour que sur chacun d'eux, la variation des valeurs de f n'excède pas ε.*

► Sur un intervalle qui n'est pas un segment, une fonction, même continue, n'atteint pas nécessairement un maximum : par exemple, la fonction $x \mapsto x^2$ sur $[0, 1[$ n'atteint pas son maximum sur $[0, 1[$ (figure 1).

► Si une fonction n'est pas continue en un point d'un segment, elle n'a pas forcément non plus de maximum sur ce segment : par exemple, en désignant par $E(x)$ la partie entière du nombre x, la fonction $x \mapsto x - E(x)$ n'atteint pas son maximum sur $[0,1]$ (figure 2).

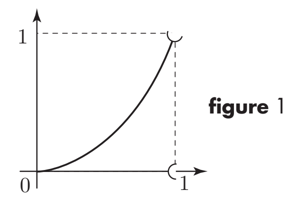


figure 1

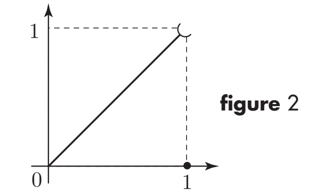


figure 2

La propriété (b) s'appelle la *continuité uniforme*. Elle exprime que si l'on se donne $\varepsilon > 0$, alors il existe $\alpha > 0$ tel que, pour tous x et y dans $[a,b]$ vérifiant $|x-y| \leqslant \alpha$, on a $|f(x) - f(y)| \leqslant \varepsilon$.
Une fonction continue sur un intervalle I qui n'est pas un segment n'a pas forcément la propriété de continuité uniforme.

Par exemple, pour la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^2$, les valeurs $x_n = n$ et $y_n = x_n + 1/n$ sont très proches lorsque n est assez grand, mais les valeurs $f(x_n) = n^2$ et $f(y_n) = (n+1/n)^2$ diffèrent toujours de $f(y_n) - f(x_n) = (n+1/n)^2 - n^2 = 2 + (1/n)^2 > 2$.

5. L'intégrale

5.1 Définition de l'intégrale

Il s'agit de mesurer l'aire comprise entre le graphe d'une fonction et l'axe des x, sur un segment donné $I = [a,b]$.
Pour cette question, les fonctions les plus simples sont les « fonctions en escalier » : elles sont constantes sur des sous-intervalles formant une partition de I (figure 1).
Si la fonction v prend les valeurs $v_1, v_2, \ldots, v_p$ sur ces intervalles, l'intégrale de v est par définition le nombre

$$I(v) = v_1\ell_1 + v_2\ell_2 + \cdots + v_p\ell_p$$

où $\ell_1, \ldots, \ell_p$ sont les longueurs des intervalles.
Dans l'intégrale, chaque aire de rectangle est comptée avec le signe de v_i : positivement si v_i est positif, négativement sinon. L'intégrale est donc une aire algébrique.

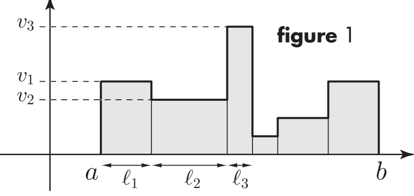


figure 1

Pour définir l'intégrale d'une fonction f plus générale, on l'encadre par des fonctions en escalier u et v prenant des valeurs $u_1, \ldots, u_p$ et $v_1, \ldots, v_p$ sur des sous-intervalles : sur la figure 2, l'aire $\mathscr{A}$ sous le graphe de f satisfait ainsi

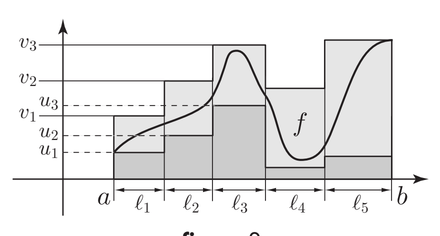


figure 2

l'encadrement

$I(u) = u_1\ell_1 + u_2\ell_2 + u_3\ell_3 + u_4\ell_4 + u_5\ell_5 \leqslant \mathscr{A} \leqslant v_1\ell_1 + v_2\ell_2 + v_3\ell_3 + v_4\ell_4 + v_5\ell_5 = I(v)$

Si l'on peut construire des encadrements de plus en plus précis, de manière que $I(v) - I(u)$ tende vers 0, on prendra comme définition de l'intégrale de f la limite commune des nombres $I(u)$ et $I(v)$.

Définition

Soit f une fonction à valeurs réelles définie sur le segment $[a, b]$. On dit que f est *intégrable* si pour tout nombre $\varepsilon > 0$, il existe des fonctions en escalier u et v telles que $u(x) \leqslant f(x) \leqslant v(x)$ pour tout $x \in [a, b]$ et $I(v) - I(u) \leqslant \varepsilon$.

Toute fonction intégrable est à la fois majorée et minorée.

Théorème. *Une fonction monotone sur un segment est intégrable.*

Démonstration. Partageons le segment $[a, b]$ en n parties égales de longueur $\ell = (b-a)/n$:

$$I_1 = [a, a_1]\ ,\quad I_2 = [a_1, a_2]\ ,\quad \ldots,\quad I_n = [a_{n-1}, b]\,,$$

où $a_1 = a + \ell, a_2 = a + 2\ell, \ldots, a_{n-1} = a + (n-1)\ell$; posons de plus $a_0 = a$ et $a_n = b$. Si f est une fonction croissante sur $[a, b]$, définissons des fonctions en escalier u et v comme suit :

$$u(x) = f(a_i) \quad \text{et} \quad v(x) = f(a_{i+1})\ ,\quad \text{si } x \in [a_i, a_{i+1}[\text{ et } 0 \leqslant i \leqslant n-1$$

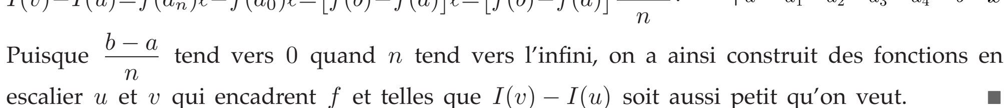
Pour tout x dans l'intervalle $[a_i, a_{i+1}[$, on a $u(x) = f(a_i) \leqslant f(x) \leqslant f(a_{i+1}) = v(x)$, car f est croissante : les fonctions u et v encadrent donc f. Par définition, les intégrales de u et de v sont

$$I(u) = f(a_0)\ell + f(a_1)\ell + \cdots f(a_{n-1})\ell$$

$$I(v) = f(a_1)\ell + \ldots + f(a_{n-1})\ell + f(a_n)\ell\,,$$

donc

$$I(v) - I(u) = f(a_n)\ell - f(a_0)\ell = \big[f(b) - f(a)\big]\ell = \big[f(b) - f(a)\big]\,\frac{b-a}{n}.$$



Puisque $\dfrac{b-a}{n}$ tend vers 0 quand n tend vers l'infini, on a ainsi construit des fonctions en escalier u et v qui encadrent f et telles que $I(v) - I(u)$ soit aussi petit qu'on veut. ■

Théorème. *Une fonction continue sur un segment est intégrable.*

Démonstration. En utilisant la propriété (b) énoncée page 283, nous allons construire pour tout nombre $\varepsilon > 0$, des fonctions en escalier u et v telles que $u(x) \leqslant f \leqslant v(x)$ et $I(v) - I(u) \leqslant \varepsilon$.

- On partage le segment $[a, b]$ en n sous-segments $I_1, \ldots, I_n$ sur chacun desquels f varie d'au plus $\varepsilon/(b-a)$;
- Puisque f est continue sur le segment I_k, elle a un maximum sur I_k : on prend ce maximum comme valeur de $v(x)$ pour $x \in I_k$ (extrémité droite exclue). De même, on définit u en lui donnant sur I_k la valeur minimum de f sur I_k.

Pour tout $x \in I_k$, on a $u(x) \leqslant f(x) \leqslant v(x)$ et comme les valeurs $f(x)$ varient d'au plus $\varepsilon/(b-a)$ sur I_k, on a $v(x) - u(x) \leqslant \varepsilon/(b-a)$ quel que soit x. En notant $\ell_1, \ldots, \ell_n$ les longueurs des

segments $I_1, \ldots, I_n$, il vient

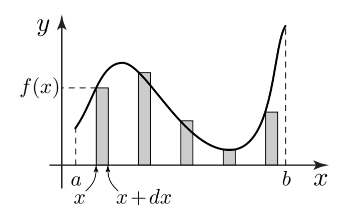
$$I(v) - I(u) \leqslant \frac{\varepsilon}{b-a}\ell_1 + \cdots + \frac{\varepsilon}{b-a}\ell_n = \frac{\varepsilon}{b-a}(\ell_1 + \cdots + \ell_n) = \frac{\varepsilon}{b-a}(b-a) = \varepsilon\,.$$

■

Existence de l'intégrale. *Si une fonction $f : [a,b] \to \mathbb{R}$ est intégrable, il existe un unique nombre $I(f)$, appelé l'intégrale de f, tel que $I(u) \leqslant I(f) \leqslant I(v)$ pour toutes fonctions en escalier u et v vérifiant $u(x) \leqslant f(x) \leqslant v(x)$. L'intégrale de f se note $\int_a^b f(t)\,dt$.*

Nous admettons l'existence de l'intégrale. On peut aussi montrer que les fonctions en escalier obtenues en subdivisant $[a,b]$ en sous-intervalles de plus en plus petits ont leur intégrale qui tend vers l'intégrale de f (pour une fonction f monotone ou continue, cela ressort des démonstrations ci-dessus.) Intuitivement, l'intégrale de f sur $[a,b]$ est bien l'aire algébrique comprise entre l'axe des abscisses et le graphe de f.

Le signe $\int$ évoque la lettre S, initiale de « somme » et la notation $\int_a^b f(t)\,dt$ suggère qu'il faut penser l'intégrale comme la limite d'une somme de quantités infiniment petites $\delta y = f(x)\delta x$: on imagine que δy est l'aire d'un rectangle de hauteur $f(x)$ et de base infiniment petite δx.



Raisonnement par infiniments petits. En Physique, on raisonne souvent en décomposant une quantité en une « somme d'infiniments petits » : cela conduit finalement à exprimer cette quantité sous la forme d'une intégrale.

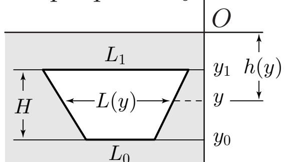
Exemple. La pression exercée sur un objet immergé dans un fluide à la profondeur h est $P = \rho h$, où ρ est la masse volumique du fluide. Cherchons la force exercée sur une plaque verticale immergée.

Choisissons un axe vertical Oy et considérons dans la plaque une tranche horizontale de hauteur infiniment petite δy située à la profondeur h ; sa longueur L est fonction de y. Sur cette tranche d'aire $L\delta y$, le fluide exerce une force horizontale d'intensité $PL\delta y = \rho h L \delta y$. On en déduit que la force exercée sur la plaque a pour intensité

$$F = \int_{y_0}^{y_1} \rho h(y) L(y)\,dy$$

où y_0 et y_1 sont les valeurs de y à la base et au sommet de la plaque.

Supposons par exemple qu'un canal est fermé par une vanne coulissante en forme de trapèze. Dirigeons l'axe Oy vers le haut et prenons son origine à la surface de l'eau. La profondeur est donc $h(y) = -y$. Notons L_0 la longueur de la base inférieure de la plaque, L_1 la longueur du coté supérieur et H la hauteur. Entre y_0 et y_1, $L(y)$ varie proportionnellement :



$$\frac{L(y) - L_0}{L_1 - L_0} = \frac{y - y_0}{y_1 - y_0} = \frac{y - y_0}{H}$$

d'où $L(y) = L_0 + \frac{L_1 - L_0}{H}(y - y_0) = \frac{L_1 - L_0}{H} y + L_0 - \frac{L_1 - L_0}{H} y_0$. Il vient, puisque $\rho = 1$:

$$F = \int_{y_0}^{y_1} -yL(y)\,dy = \int_{y_0}^{y_1} \left[-\frac{L_1 - L_0}{H} y^2 - \frac{L_0 y_1 - L_1 y_0}{H} y \right] dy$$

5.2 Propriétés de l'intégrale

Elles résultent des propriétés de l'intégrale d'une fonction en escalier.

Supposons que u est une fonction en escalier, constante sur les intervalles d'extrémités $a = a_0 < a_1 < \cdots < a_{n-1} < a_n = b$.
Pour tout nombre réel λ, la fonction $\lambda u : x \mapsto \lambda u(x)$ est constante sur ces mêmes intervalles, donc est en escalier, et l'on a $I(\lambda u) = \lambda I(u)$.
Soit u' une fonction en escalier constante sur les intervalles d'extrémités $a < a'_1 < \cdots < a'_{p-1} < b$.
En mettant ensemble les points a_i et a'_i, on définit un partage de $[a, b]$ en intervalles sur lesquels u et u' sont constantes : les fonctions $u + u'$ et uu' sont alors constantes sur ces intervalles. Remarquons que l'on a $I(u + u') = I(u) + I(u')$ et que si u est une fonction en escalier positive ou nulle, alors $I(u)$ est un nombre positif ou nul.
En passant à la limite quand la longueur des intervalles tend vers 0, on obtient les propriétés suivantes.

a) *Si f est intégrable sur $[a, b]$, alors pour tout nombre λ, la fonction $x \mapsto \lambda f(x)$ est intégrable et l'on a $\int_a^b \lambda f(t)\,dt = \lambda \int_a^b f(t)\,dt$.*

b) *Si f et g sont intégrables sur $[a, b]$, alors la fonction $x \mapsto f(x) + g(x)$ est intégrable et l'on a $\int_a^b \big[f(t) + g(t)\big]\,dt = \int_a^b f(t)\,dt + \int_a^b g(t)\,dt$.*

c) *Si $f(x) \geqslant 0$ pour tout x, alors $\int_a^b f(t)\,dt \geqslant 0$.*

d) *Si f et g sont intégrables sur $[a, b]$, leur produit $x \mapsto f(x)g(x)$ est intégrable.*

Les deux premières propriétés signifient que l'ensemble des fonctions intégrables sur $[a, b]$ est un espace vectoriel et que l'application $f \mapsto \int_a^b f(t)\,dt$ est une application linéaire.
La propriété (c) affirme que l'intégrale d'une fonction positive ou nulle est positive ou nulle. Si f et g sont intégrables et si $f \geqslant g$, alors la fonction $f - g$ est positive ou nulle, donc son intégrale aussi :

$$\textit{si } f(x) \geqslant g(x) \textit{ pour tout } x \in [a, b], \textit{ alors } \int_a^b f(t)\,dt \geqslant \int_a^b g(t)\,dt.$$

Si f est intégrable sur $[a, b]$, elle est intégrable sur tout segment $[c, d]$ inclus dans $[a, b]$. On pose :

$$\int_d^c f(t)\,dt = -\int_c^d f(t)\,dt \quad \text{et} \quad \int_c^c f(t)\,dt = 0 \text{ pour tous nombres } c \text{ et } d \text{ dans } [a, b].$$

Pour tous nombres x, y, z dans $[a, b]$, on a alors l'égalité :

$$\int_x^y f(t)\,dt + \int_y^z f(t)\,dt = \int_x^z f(t)\,dt \qquad \textit{(relation de Chasles)}$$

Inégalité fondamentale. *Si f est une fonction intégrable sur $[a, b]$, alors*

$$\left|\int_a^b f(t)\,dt\right| \leqslant \int_a^b |f(t)|\,dt$$

Démonstration. Puisqu'on a $-|f(x)| \leqslant f(x) \leqslant |f(x)|$ pour tout x, il vient en effet $-\int_a^b |f(t)|\,dt = \int_a^b -|f(t)|\,dt \leqslant \int_a^b f(t)\,dt \leqslant \int_a^b |f(t)|\,dt$. ■

Proposition. *Soit f une fonction continue et positive ou nulle sur le segment $[a, b]$. Si $\int_a^b f(t)\,dt = 0$, alors $f(x) = 0$ quel que soit $x \in [a, b]$.*

Démonstration. Raisonnons par l'absurde en supposant que f prend en un point $c \in [a, b]$ une valeur $f(c) > 0$. Puisque f est continue en c, il y a un intervalle autour de c où les valeurs de f sont supérieures à un nombre $m > 0$. En appelant u et v les extrémités de cet intervalle, on a $\int_a^b f(t)\,dt \geqslant \int_u^v f(t)\,dt$, car f est positive ou nulle, d'où $\int_a^b f(t)\,dt \geqslant (v-u)m > 0$, ce qui est une contradiction. ■

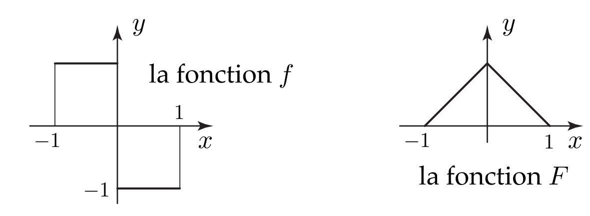
Intégrale fonction de la borne supérieure

Proposition. *Si f est une fonction intégrable sur $[a, b]$, alors la fonction $x \mapsto \int_a^x f(t)\,dt$ est continue sur $[a, b]$.*

Démonstration. Posons en effet $F(x) = \int_a^x f(t)\,dt$. Si $x_0 \in [a, b]$, alors d'après la relation de Chasles, on a $\int_a^x f(t)\,dt = \int_a^{x_0} f(t)\,dt + \int_{x_0}^x f(t)\,dt$, ou encore $F(x) = F(x_0) + \int_{x_0}^x f(t)\,dt$. Puisque f est majorée et minorée, il existe un nombre M tel que $|f(x)| \leqslant M$ pour tout x. D'après l'inégalité fondamentale, il vient $\left|F(x) - F(x_0)\right| \leqslant \left|\int_{x_0}^x |f(t)|\,dt\right| \leqslant \left|\int_{x_0}^x M\,dt\right| = M|x - x_0|$. Quand x tend vers x_0, $|x - x_0|$ tend vers 0 et donc $F(x)$ tend vers $F(x_0)$. ■

Exemple. Prenons la fonction en escalier définie sur $[-1,1]$ par $f(x)=1$ si $-1 \leqslant x \leqslant 0$ et $f(x) = -1$ si $0 < x \leqslant 1$. Posons $F(x) = \int_{-1}^x f(t)\,dt$.

- Si $x \in [-1, 0]$: alors $f(t) = 1$ sur l'intervalle $[-1, x]$, donc $F(x) = \int_{-1}^x 1\,dt = x+1$.
- Si $x \in]0, 1]$: alors $f(t) = -1$ sur l'intervalle $]0, x]$, donc $\int_0^x f(t)\,dt = \int_0^x (-1)\,dt = -x$ et $F(x) = \int_{-1}^0 f(t)\,dt + \int_0^x f(t)\,dt = F(0) - x = 1 - x$.



La fonction f n'est pas continue en 0, mais F est continue sur $[-1,1]$.

Voici une proposition permettant, pour des fonctions positives, d'intégrer un infiniment petit ou un équivalent.

Proposition. *Soient u et v des fonctions intégrables sur $[a,b]$ et soit $x_0 \in [a,b]$. Supposons $v(x) > 0$ pour tout $x \neq x_0$.*

- *Si $u(x) \underset{x\to x_0}{\ll} v(x)$, alors $\left(\int_{x_0}^x u(t)\,dt\right) \underset{x\to x_0}{\ll} \left(\int_{x_0}^x v(t)\,dt\right)$.*
- *Si $u(x) \underset{x\to x_0}{\sim} v(x)$, alors $\left(\int_{x_0}^x u(t)\,dt\right) \underset{x\to x_0}{\sim} \left(\int_{x_0}^x v(t)\,dt\right)$.*

Démonstration. Supposons $u(x)$ infiniment petit devant $v(x)$ quand x tend vers x_0. Alors $\varphi(x) = \dfrac{u(x)}{v(x)}$ tend vers 0 et $u(x) = v(x)\,\varphi(x)$. Étant donné un nombre $\varepsilon > 0$, on a $|\varphi(x)| \leqslant \varepsilon$ pour tout x assez proche de x_0, donc aussi $|u(x)| \leqslant \varepsilon v(x)$, car $v(x)$ est positif. Pour de tels x, il vient $\left|\int_{x_0}^x u(t)\,dt\right| \leqslant \left|\int_{x_0}^x |u(t)|\,dt\right| \leqslant \left|\int_{x_0}^x \varepsilon v(t)\,dt\right| = \varepsilon \left|\int_{x_0}^x v(t)\,dt\right|$ et le rapport $\left(\int_{x_0}^x u(t)\,dt\right) \Big/ \left(\int_{x_0}^x v(t)\,dt\right)$ est en valeur absolue inférieur à ε.

Supposons $u(x)$ équivalent à $v(x)$ quand x tend vers x_0, donc $\left[u(x)-v(x)\right] \underset{x\to x_0}{\ll} v(x)$. D'après ce qu'on vient de montrer, on a $\int_{x_0}^x \left[u(t)-v(t)\right] dt \underset{x\to x_0}{\ll} \int_{x_0}^x v(t)\,dt$, ou encore

$$\left[\int_{x_0}^x u(t)\,dt - \int_{x_0}^x v(t)\,dt\right] \underset{x\to x_0}{\ll} \int_{x_0}^x v(t)\,dt\,.$$

Ainsi, $\int_{x_0}^x u(t)\,dt$ est équivalent à $\int_{x_0}^x v(t)\,dt$ quand x tend vers x_0. ■

5.3 Moyenne d'une fonction

Définition

Si f est une fonction intégrable sur $[a,b]$, le nombre $\dfrac{1}{b-a}\displaystyle\int_a^b f(t)\,dt$ s'appelle la *moyenne de f sur $[a,b]$*.

Proposition. *Soit f une fonction continue sur $[a,b]$ et soit w une fonction intégrable à valeurs positives ou nulles et d'intégrale strictement positive. Alors il existe un nombre*

$c \in [a, b]$ tel que

$$\int_a^b w(t)f(t)\,dt = f(c)\int_a^b w(t)\,dt$$

Démonstration. Soit m le minimum de f sur $[a, b]$ et M son maximum. Pour tout $t \in [a, b]$, on a $m \leqslant f(t) \leqslant M$, donc aussi $mw(t) \leqslant f(t)w(t) \leqslant Mw(t)$, car $w(t)$ est positif ou nul. En intégrant, on obtient $m\int_a^b w(t)\,dt \leqslant \int_a^b w(t)f(t)\,dt \leqslant M\int_a^b w(t)\,dt$. Le nombre $\left(\int_a^b w(t)f(t)\,dt\right) \Big/ \left(\int_a^b w(t)\,dt\right)$ est entre m et M, donc est égal à une certaine valeur $f(c)$ de la fonction f, d'après le théorème des valeurs intermédiaires. ■

En choisissant $w(t) = 1$ pour tout t, on obtient le résultat suivant.

Formule de la moyenne. *Si f est une fonction continue sur $[a,b]$, il existe $c \in [a,b]$ tel que $\int_a^b f(t)\,dt = (b-a)f(c)$.*

Pour une fonction continue, la moyenne $\dfrac{1}{b-a}\displaystyle\int_a^b f(t)\,dt$ est donc l'une des valeurs de la fonction.

Moyenne d'une fonction monotone

Supposons que f est une fonction monotone sur $[a,b]$ et partageons $[a,b]$ en n segments de même longueur $h=(b-a)/n$, donc d'extrémités $a=a_0, a_1=a+h, a_2=a+2h, \ldots, a_{n-1}=a+(n-1)h, a_n=b$. Nous avons montré page 285 que $\dfrac{b-a}{n}\big[f(a_0)+f(a_1)+\cdots+f(a_{n-1})\big]$ tend vers $\int_a^b f(t)\,dt$ quand n tend vers l'infini.

Quand n tend vers l'infini, $\dfrac{1}{n}\big[f(a_0)+f(a_1)+\cdots+f(a_{n-1})\big]$ tend vers $\dfrac{1}{b-a}\displaystyle\int_a^b f(t)\,dt$.

Si l'on fait la moyenne des valeurs de f en des points régulièrement répartis dans $[a, b]$, on obtient des nombres qui tendent vers la moyenne de f quand le nombre de points tend vers l'infini.

5.4 Primitives

Si une fonction f est intégrable, la fonction $F(x) = \int_a^x f(t)\,dt$ est continue (proposition page 288). Nous allons voir que si f est continue, alors $F(x)$ est dérivable.
Rappelons que, par définition, une *primitive* de f est une fonction Φ dérivable telle que $\Phi' = f$.

Théorème fondamental. *Si f est une fonction continue sur un intervalle I et si $a \in I$, alors la fonction $F(x) = \int_a^x f(t)\,dt$ est une primitive de f : on a $F'(x) = f(x)$ pour tout $x \in I$.*

Démonstration. Donnons-nous $x_0 \in I$ et posons $y_0 = f(x_0)$. Puisque $f(x) = y_0 + \big(f(x) - y_0\big)$, on a

$$F(x) - F(x_0) = \int_{x_0}^{x} f(t)\,dt = \int_{x_0}^{x} y_0\,dt + \int_{x_0}^{x} \big(f(t) - y_0\big)dt = y_0(x - x_0) + \int_{x_0}^{x} \big(f(t) - y_0\big)dt$$

Posons $u(x) = \int_{x_0}^{x} \big(f(t) - y_0\big)dt$. Étant donné un nombre $\varepsilon > 0$, on sait que pour tout t assez proche de x_0, on a $|f(t) - y_0| \leqslant \varepsilon$, donc pour x proche de x_0, il vient

$$|u(x)| = \left|\int_{x_0}^{x} \big(f(t) - y_0\big)dt\right| \leqslant \left|\int_{x_0}^{x} |f(t) - y_0|\,dt\right| \leqslant \left|\int_{x_0}^{x} \varepsilon\,dt\right| = |x - x_0|\varepsilon$$

Ainsi $u(x)$ est infiniment petit devant $|x - x_0|$ quand x tend vers x_0. Puisque $F(x) = F(x_0) + y_0(x - x_0) + u(x)$, cela signifie d'après la proposition page 271, que l'on a $F'(x_0) = y_0$. ■

Ce théorème affirme que toute fonction continue sur un intervalle possède une primitive.

Rappel. *Si F est une primitive de f sur un intervalle, les autres primitives de f sont les fonctions $F(x) + c$, où c est une constante quelconque.*

Nous verrons en effet au début du prochain chapitre que si l'on a $F'(x) - G'(x) = 0$ en tout point x d'un intervalle, alors la fonction $F - G$ est constante sur cet intervalle.

Exemple. Reprenons la fonction F de l'exemple page 288 et calculons $U(x) = \int_{-1}^{x} F(t)\,dt$ pour $x \in [-1, 1]$.

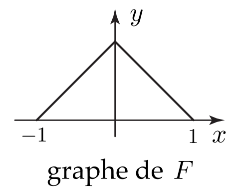
Puisque F est continue sur $[-1, 1]$, on sait que la fonction U est une primitive de F : on a $U'(x) = F(x)$ pour tout $x \in\,]-1, 1[$.

Pour $x \in [-1, 0]$, on a $F(x) = x + 1$, donc $U(x) = x^2/2 + x + c$, où c est une constante. Par définition de U, on a $U(-1) = 0$, donc $1/2 - 1 + c = 0$ et $c = 1/2$:

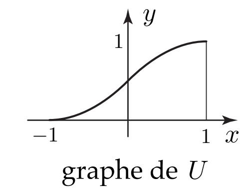
$$U(x) = \frac{x^2}{2} + x + \frac{1}{2}, \text{ pour tout } x \in [-1, 0].$$

Supposons $x \in [0, 1]$. Alors $U(x) = U(0) + \int_0^x F(t)\,dt = 1/2 + \int_0^x (-t + 1)\,dt$. L'intégrale vaut $-x^2/2 + x + d$ et elle est nulle pour $x = 0$, donc $d = 0$. Par suite

$$U(x) = -\frac{x^2}{2} + x + \frac{1}{2}, \text{ pour tout } x \in [0, 1].$$



graphe de F



graphe de U

Nous voyons que lorsqu'on intègre une fonction, on la rend plus régulière : la fonction F n'est pas dérivable en 0, mais la fonction U est dérivable.

Intégrale d'une fonction paire ou impaire. Soit f une fonction continue sur $[-a,a]$. Posons $F(x) = \int_0^x f(t)\,dt$, pour tout $x \in [-a,a]$.

► Si la fonction f est paire, alors

$$\int_0^{-x} f(t)\,dt = -\int_0^x f(t)\,dt \quad \text{et} \quad \int_{-x}^x f(t)\,dt = 2\int_0^x f(t)\,dt \ , \quad \text{pour tout } x \in [-a,a].$$

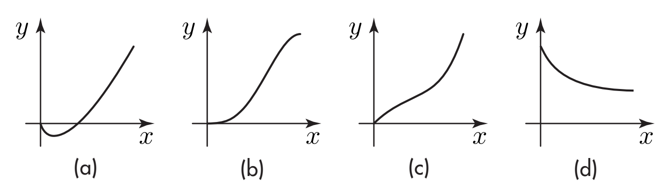
► Si f est impaire, alors

$$\int_0^{-x} f(t)\,dt = \int_0^x f(t)\,dt \quad \text{et} \quad \int_{-x}^x f(t)\,dt = 0 \ , \quad \text{pour tout } x \in [-a,a].$$

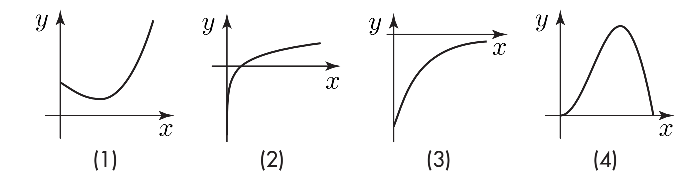
Posons en effet $F(x)=\int_0^x f(t)\,dt$. La dérivée de $F(-x)$ est $-F'(-x)=-f(-x)$, car $F'=f$. Si f est paire, alors $\frac{d}{dx}\big[F(-x)\big] = -f(-x) = -f(x) = \frac{d}{dx}\big[-F(x)\big]$. Les fonctions $F(-x)$ et $-F(x)$ ont la même dérivée et prennent en $x=0$ la même valeur 0, donc elles sont égales. D'après la relation de Chasles, il vient alors $\int_{-x}^x f(t)\,dt = \int_{-x}^0 f(t)\,dt + \int_0^x f(t)\,dt = -F(-x) + F(x) = 2F(x)$. On raisonne de même si f est impaire : dans ce cas, les fonctions $F(-x)$ et $F(x)$ ont même dérivée et coïncident en 0, donc elles sont égales.

Exercices

@ **1.** À chaque graphe d'une fonction f ci-dessous, associer le graphe de la dérivée f'.
graphe de la fonction f :

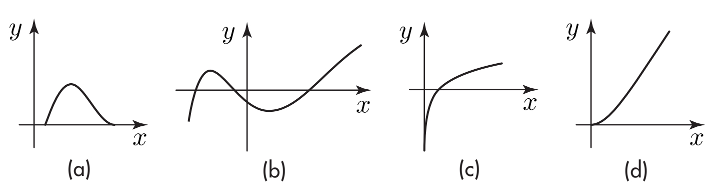


graphe de la dérivée f' :

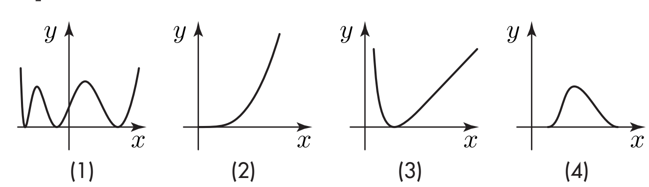


@ **2.** À chaque graphe d'une fonction f ci-dessous, associer le graphe du carré de f et de la racine carrée de f.

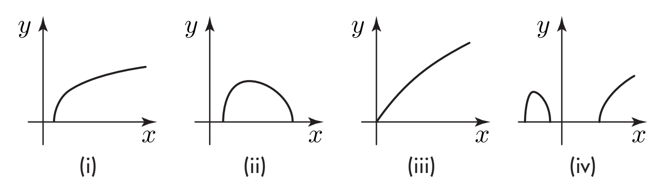
graphe de $f(x)$:



graphe de $\left[f(x)\right]^2$:



graphe de $\sqrt{f(x)}$:



3. **Application de la formule de Stirling.** Montrer que le coefficient binomial $\binom{2n}{n}$ est équivalent à $\dfrac{1}{\sqrt{\pi}}\dfrac{4^n}{\sqrt{n}}$ quand n tend vers l'infini.

@ 4. **Étude d'un point fixe**

a) Représenter sur un même dessin les graphes des fonctions x, $e^{-(1/2)x}$, e^{-x} et e^{-2x}. Étant donné un nombre réel $a \geqslant 0$, quel est le signe de $e^{-ax}-x$ pour x très grand ? Quel est le signe pour $x < 0$? Montrer que, pour tout $a \geqslant 0$, la fonction $f(x) = e^{-ax}$ a un unique point fixe ; on note $s(a)$ ce point fixe.

b) Que vaut $s(0)$? Montrer que l'on a $e^{-ax}-x > 0$ si $0 \leqslant x < s(a)$ et $e^{-ax}-x < 0$ si $x > s(a)$. Montrer que si $0 < a < b$, alors $e^{-as(b)}-s(b) > 0$. En déduire que la fonction $a \mapsto s(a)$ est décroissante. Montrer que $s(a)$ tend vers 0 quand a tend vers $+\infty$ (raisonner par l'absurde en supposant que la limite est un nombre $\ell > 0$ et en utilisant l'égalité $as(a) = -\ln s(a)$).

c) Montrer que la fonction $t \mapsto ts(t)$ est croissante et que l'on a $s(t) \underset{t\to+\infty}{\gg} 1/t$ (utiliser le résultat précédent). Dessiner l'allure de la fonction $t \mapsto s(t)$.

d) Admettons que la fonction s est dérivable. En utilisant l'égalité $s(t) = e^{-ts(t)}$, montrer que la fonction $s(t)$ est solution de l'équation différentielle $y' = -\dfrac{y^2}{1+ty}$.

@ **5. Somme de plusieurs fonctions périodiques.** On dit qu'une fonction f est périodique (de période $T \neq 0$) si l'on a $f(x+T) = f(x)$ quel que soit $x \in \mathbb{R}$.

a) Soit f une fonction périodique de période T et dérivable. Montrer que f' est périodique de période T. Montrer que la fonction $1+\cos x$ est périodique, mais qu'elle n'a aucune primitive périodique.

b) Posons $f(x) = a\sin\alpha x + b\sin\beta x$, où a, b, α, β sont des nombres réels tous non nuls. Nous allons voir que f ne peut être périodique que si β/α est un nombre rationnel.

(i) Supposons que f est périodique de période T. En considérant f'', montrer que $\begin{cases} a\sin\alpha T + b\sin\beta T &= 0 \\ a\alpha^2\sin\alpha T + b\beta^2\sin\beta T &= 0 \end{cases}$. Montrer que si $\alpha \neq \pm\beta$, alors $\sin\alpha T = \sin\beta T = 0$. En déduire qu'il existe des entiers n et k tels que $\beta/\alpha = n/k$.

(ii) Montrer que si $\beta/\alpha = n/k$, avec n et k entiers, alors f a pour période $\dfrac{2\pi k}{\alpha}$.

c) Montrer que la fonction $\sin\left(\dfrac{x}{2}+\varphi\right) + \sin(\sqrt{3}x)$ n'est pas périodique (en procédant comme ci-dessus, montrer que si cette fonction avait une période T, alors on aurait $\sin\left(\dfrac{T}{2}+\varphi\right) = \sin\varphi$, puis en déduire une contradiction).

d) Soit q un entier positif. La fonction $2\sin 3x + \sin\dfrac{2x}{3} - \cos\dfrac{x}{q}$ est périodique : quelle est sa période ? (la réponse n'est pas la même selon que q est multiple de 3, ou non).

6. À la distance r du centre d'une artère de rayon R, la vitesse du sang est $v = c(R^2 - r^2)$, où c est une constante. Quand on administre une substance par voie intraveineuse, l'artère se dilate légèrement, à un rythme constant $\rho = \dfrac{dR}{dt}$. De combien varie la vitesse du sang à la distance r du centre ? ($c = 3{,}2.10^3$ cm^{-1}s^{-1}, $R = 0{,}5$ cm et $\rho = 5.10^{-3}$ cm s^{-1})

@ **7. Utilisation de la différentielle logarithmique.** Un ballon sphérique en matière élastique, rempli d'air, subit simultanément une variation relative de pression de 5% et une variation relative de température de 2%. On suppose que l'air suit la loi $PV^\gamma/T =$ constante, où P est la pression, V le volume, T la température absolue et où $\gamma = 1{,}4$. Quelle est la variation relative du diamètre du ballon ?

8. Au cours d'un trajet sur autoroute, un automobiliste a parcouru 200 km en deux heures (avec des arrêts éventuels). On mesure le temps t en heures et l'on note $f(t)$ le nombre de kilomètres parcourus depuis le départ. Posons $g(t) = f(t+1) - f(t)$ pour $0 \leqslant t \leqslant 1$.

a) Montrer que $g(t)$ est positif ou nul et que $g(0) + g(1) = 200$. En déduire que la fonction $g(t) - 100$ ne garde pas un signe constant.

b) Montrer qu'il existe une durée d'une heure pendant laquelle l'automobiliste a parcouru exactement 100 km.

@ 9. Comparaison d'ordres de grandeur

a) Comparer les ordres de grandeur des fonctions $\sqrt{x}$ et $(\ln x)^3$ quand x tend vers $+\infty$ et en déduire la partie principale de $\sqrt{x} + (\ln x)^3$.

b) Comparer les ordres de grandeur des fonctions xe^{-x}, e^{-x}/x et e^{-x^2} quand x tend vers $+\infty$.

c) Montrer que $\int_0^x t[1 + \cos t]^\alpha \, dt$ est équivalent à $2^{\alpha-1}x^2$ quand x tend vers 0.

d) Montrer que $\dfrac{\ln x}{x^4+3x^2-2x^3-4x+2}$ a pour partie principale $\dfrac{1}{3(x-1)}$ quand x tend vers 1.

10. a) Justifier les formules suivantes :

(i) $\displaystyle\int_{-a}^{a} \left(e^t - e^{-t}\right)(\sin t)^2 \, dt = 0$ **(ii)** $\displaystyle\int_0^{\pi/4} \left[1 + (\cos t)^4\right] e^t \, dt \geqslant (5/4)(e^{\pi/4} - 1)$

b) Dessiner le graphe de la fonction $\sqrt{1-x^3}$ entre 0 et 1. En déduire que si $0 < a < b < 1$, alors $\displaystyle\int_a^b \sqrt{1-t^3} \, dt \geqslant \frac{b-a}{2}\left[\sqrt{1-a^3} + \sqrt{1-b^3}\right]$.

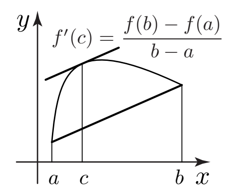
Chapitre 10

Utilisation de la dérivée et de l'intégrale

1. Étude des variations d'une fonction

La dérivée $f'(a)$ d'une fonction f en un point a est le taux de proportionnalité entre les infiniments petits $f(x)-f(a)$ et $x-a$, quand x tend vers a : la dérivée en a ne fait intervenir que les valeurs $f(x)$ pour x voisin de a.
Nous allons voir que si f a une dérivée en tout point d'un intervalle I, les nombres $f'(x)$ donnent un contrôle sur tous les taux d'accroissement $\frac{f(b)-f(a)}{b-a}$, quels que soient a et b dans I.

Théorème des accroissements finis. *Soit f une fonction dérivable sur un intervalle I. Pour tous nombres a et b dans I, il existe un nombre c strictement compris entre a et b tel que $f(b)-f(a)=(b-a)f'(c)$.*



La tangente en c est parallèle à la corde

Démonstration. Posons $k = \dfrac{f(b)-f(a)}{b-a}$ et $\varphi(x) = k(x-a) + f(a)$. Les fonctions f et φ sont continues, donc aussi leur différence $u(x) = f(x) - \varphi(x)$. D'après les propriétés des fonctions continues sur un segment, il y a un nombre c entre a et b où $u(x)$ atteint son maximum (page 283). On sait qu'en ce point c, la tangente au graphe de u est horizontale, donc $u'(c) = f'(c) - \varphi'(c) = 0$. Puisque $\varphi'(x) = k$, il vient $f'(c) = k$. ■

Une première conséquence du théorème, c'est qu'au moyen de la dérivée, on peut caractériser les fonctions constantes, les fonctions croissantes et les fonctions décroissantes.

Caractérisation des fonctions constantes. Si $f'(x)=0$ pour tout $x \in I$, alors f est constante.

Caractérisation des fonctions monotones

- Si $f'(x) \geqslant 0$ pour tout $x \in I$, alors f est croissante sur I. Si $f'(x) > 0$ sauf peut-être pour un nombre fini de valeurs de x, alors f est strictement croissante sur I.
- De même, si $f'(x) \leqslant 0$ pour tout $x \in I$, f est décroissante sur I.

Application aux primitives

Soit f une fonction continue sur un intervalle I et soit $x_0 \in I$.

► *Si F est une primitive de f, alors $F(x) - F(x_0) = \int_{x_0}^{x} f(t)\,dt$.*

► *Toute primitive de f s'écrit $F(x) = \int_{x_0}^{x} f(t)\,dt + c$, où c est une constante.*

On sait que la fonction $U(x) = \int_{x_0}^{x} f(t)\,dt$ est une primitive de f telle que $U(x_0) = 0$. Si F est une autre primitive de f, alors $U' = f = F'$, $U' - F' = 0$, donc la fonction $U - F$ est constante. Comme cette fonction prend en x_0 la valeur $U(x_0) - F(x_0) = -F(x_0)$, on en déduit $U(x) = F(x) - F(x_0)$ pour tout $x \in I$.

Notation. On notera simplement $\int f(t)\,dt$ une primitive de f. On écrit alors par exemple $\sin x = \int \cos t\,dt$: c'est une égalité de fonctions à constante près.

Les primitives suivantes s'obtiennent par dérivation, en vérifiant simplement que dans chaque cas, la dérivée du second membre est égale à la fonction sous le signe intégrale.

Primitives usuelles

$$\int t^{\alpha}\,dt = \frac{1}{\alpha+1} x^{\alpha+1}\,, \text{ si } \alpha \neq -1 \qquad \int \frac{dt}{at+b} = \frac{1}{a} \ln|ax+b|$$

$$\int e^{at}\,dt = \frac{1}{a} e^{ax} \qquad \int \ln t\,dt = x \ln x - x$$

$$\int \cos(at)\,dt = \frac{1}{a}\sin(ax) \qquad \int \sin(at)\,dt = -\frac{1}{a}\cos(ax)$$

$$\int \frac{dt}{\cos^2 t} = \tan x \qquad \int \frac{dt}{\sin^2 t} = -\frac{1}{\tan x}$$

$$\int \frac{dt}{a^2+t^2} = \frac{1}{a} \operatorname{Arc}\tan\frac{x}{a} \qquad \int \frac{dt}{a^2-t^2} = \frac{1}{2a} \ln\left|\frac{a+x}{a-x}\right|$$

$$\int \frac{dt}{\sqrt{a^2+t^2}} = \ln\bigl(x + \sqrt{a^2+x^2}\bigr) \qquad \int \frac{dt}{\sqrt{a^2-t^2}} = \operatorname{Arc}\sin\frac{x}{a}$$

$$\int \frac{dt}{\sqrt{t^2-a^2}} = \ln\bigl|x + \sqrt{x^2-a^2}\bigr|$$

L'inégalité des accroissements finis

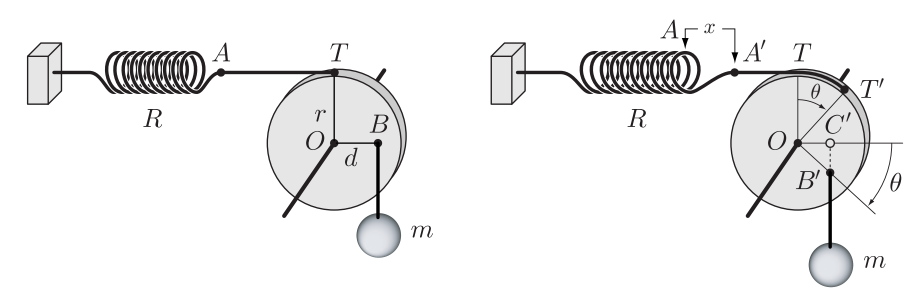
Soit f une fonction dérivable sur un intervalle I. Appliquons le théorème des accroissements finis entre des nombres x et y de I et prenons les valeurs absolues : on obtient $\bigl|f(x)-f(y)\bigr| = |x-y|\bigl|f'(z)\bigr|$, où z est un certain nombre compris entre x et y.

Proposition. *Supposons qu'on a la majoration $|f'(t)| \leqslant M$ pour tout $t \in I$. Alors pour tous nombres x et y dans I, on a $\bigl|f(x) - f(y)\bigr| \leqslant M|x-y|$.*

Cette majoration très importante s'appelle *l'inégalité des accroissements finis*.
Si l'on connaît un majorant de la fonction $t \mapsto |f'(t)|$ sur un segment, l'inégalité

des accroissements finis permet de calculer un encadrement pour les valeurs de la fonction f sur ce segment.

Exemple. Le dispositif ci-dessous montre un disque mobile autour d'un axe horizontal en O. Le point T, à la verticale de O, est relié à un ressort R par un fil AT. Au point B situé à l'horizontale de O, on laisse pendre une masse m. La roue tourne alors d'un angle θ, le ressort s'allonge de $AA'=x$ et le fil s'enroule le long de l'arc $\overset{\frown}{TT'}$.



Notons r le rayon de la roue et d la distance OB. À l'équilibre, le point d'attache du poids est en B' et le vecteur $\overrightarrow{OB'}$ fait l'angle θ avec l'horizontale.
Le poids $\vec{P}$ appliqué en B' a pour valeur mg et le ressort exerce en T une force de rappel $\vec{F}$ horizontale d'intensité kx, où k est le coefficient de dureté du ressort. À l'équilibre, les moments de $\vec{P}$ et $\vec{F}$ ont la même valeur numérique :

- le moment de $\vec{F}$ est $F \times OT = kx \times r$
- le moment de $\vec{P}$ est $P \times OC' = P \times OB' \cos\theta = mg \times d\cos\theta$.

La distance x est égale à la longueur de l'arc $\overset{\frown}{TT'}$, donc $x = r\theta$. La condition d'équilibre s'écrit $kxr = kr^2\theta = mgd\cos\theta$, c'est-à-dire

(1) $$\theta = \frac{mgd}{kr^2}\cos\theta$$

Puisqu'il existe évidemment une position d'équilibre, l'équation (1) a une solution θ_e. On peut le démontrer en introduisant la fonction continue $u(\theta) = \frac{mgd}{kr^2}\cos\theta - \theta$. La valeur $u(0) = mgd/kr^2$ est positive et $u(\pi/2) = -\pi/2 < 0$: d'après le théorème des valeurs intermédiaires, l'équation (1) a au moins une solution entre 0 et $\pi/2$. La fonction $\theta \mapsto \cos\theta$ étant strictement décroissante sur $[0, \pi/2]$, u est strictement décroissante entre 0 et $\pi/2$, donc la solution θ_e est unique.
Posons $K = \frac{mgd}{kr^2}$ et $f(\theta) = K\cos\theta$. Puisque $f(\theta_e) = \theta_e$, la solution θ_e est un point fixe de la fonction f. On a $f'(\theta) = -K\sin\theta$, $\left|f'(\theta)\right| \leqslant K$ et l'inégalité des accroissements finis pour la fonction f entre θ et θ_e s'écrit :

(2) $$\left|f(\theta) - f(\theta_e)\right| \leqslant K|\theta - \theta_e|$$

Supposons $K < 1$. Alors la fonction f est contractante au voisinage du point fixe θ_e (page 262) : le coefficient de contraction est K et le cône de contraction s'étend sur tout l'intervalle $[0, \pi/2]$. Dans ces conditions, on sait que le point fixe est attractif : la suite (u_n) des itérés d'un point quelconque $u_0 \in [0, \pi/2]$ a pour limite le point fixe. Pour calculer une valeur approchée de θ_e, il suffit de choisir un nombre initial u_0 entre 0 et $\pi/2$ et de calculer successivement les nombres u_n définis par $u_{n+1} = f(u_n)$: puisque u_n tend vers θ_e, on obtiendra, pour n assez grand, des valeurs approchées de la solution.

2. Développements limités

2.1 Approximation à l'ordre n

Si une fonction dérivable f vérifie $f(a) = f'(a) = 0$, alors quand x tend vers a, $f(x)$ est infiniment petit devant $x - a$. Plus généralement, pour une fonction f ayant des dérivées successives $f', f'', f^{(3)}, \ldots, f^{(n)}$, où $f^{(n)}$ désigne la dérivée n-ième de f, on a le résultat suivant.

Proposition. *Si* $f(a) = f'(a) = f''(a) = \cdots = f^{(n)}(a) = 0$, *alors* $f(x) \underset{x \to a}{\ll} (x-a)^n$.

Démonstration. Appliquons le théorème des accroissements finis entre a et x : on a $f(x) = f(x) - f(a) = (x - a)f'(x_1)$ avec x_1 entre a et x, donc

(1) $$\left|f(x)\right| \leqslant |x - a|\left|f'(x_1)\right| \quad \text{et} \quad |x_1 - a| \leqslant |x - a|$$

Appliquons le théorème pour f' entre a et x_1 : on a $f'(x_1) = f'(x_1) - f'(a) = (x_1 - a)f''(x_2)$, où x_2 est entre a et x_1, donc entre a et x, et par suite

(2) $$\left|f'(x_1)\right| \leqslant |x - a|\left|f''(x_2)\right| \quad \text{et} \quad |x_2 - a| \leqslant |x - a|$$

Poursuivons ainsi jusqu'à la dérivée $(n-2)$-ième :
on a $f^{(n-2)}(x_{n-2}) = f^{(n-2)}(x_{n-2}) - f^{(n-2)}(a) = (x_{n-2} - a)f^{(n-1)}(x_{n-1})$, où x_{n-1} est entre a et x_{n-2}, donc entre a et x, et il vient

$(n-1)$ $$\left|f^{(n-2)}(x_{n-2})\right| \leqslant |x - a|\left|f^{(n-1)}(x_{n-1})\right| \quad \text{et} \quad |x_{n-1} - a| \leqslant |x - a|$$

En mettant bout à bout ces $n-1$ inégalités, on obtient

$$\left|f(x)\right| \leqslant |x - a|\left|f'(x_1)\right| \leqslant |x - a|^2\left|f''(x_2)\right| \leqslant \cdots \leqslant |x - a|^{n-1}\left|f^{(n-1)}(x_{n-1})\right|$$

Divisons par $|x - a|^n$ (qui est positif) :

$$\left|\frac{f(x)}{(x-a)^n}\right| \leqslant \left|\frac{f^{(n-1)}(x_{n-1})}{x - a}\right| = \frac{\left|f^{(n-1)}(x_{n-1}) - f^{(n-1)}(a)\right|}{|x - a|} \leqslant \frac{\left|f^{(n-1)}(x_{n-1}) - f^{(n-1)}(a)\right|}{|x_{n-1} - a|}$$

Quand x tend vers a, x_{n-1} tend vers a et le rapport de droite tend par définition vers $\left|f^{(n)}(a)\right| = 0$. Le rapport $\dfrac{f(x)}{(x-a)^n}$ a donc pour limite 0 quand x tend vers a. ■

Définissons un polynôme P qui a les mêmes dérivées que f au point a jusqu'à l'ordre n : on sait que si $P = p_0 + p_1(x - a) + \cdots + p_n(x - a)^n$, alors $p_k = \dfrac{P^{(k)}(a)}{k!}$

pour tout entier $k \geqslant 0$ (si $k = 0$, on convient que $P^{(0)} = P$). En posant

$$P(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

on obtient donc un polynôme tel que $P(a) = f(a), P'(a) = f'(a), \ldots, P^{(n)}(a) = f^{(n)}(a)$.
Le polynôme P est le seul polynôme de degré inférieur ou égal à n ayant au point a les mêmes dérivées que f jusqu'à l'ordre n.
Pour la fonction $u(x) = f(x) - P(x)$, on aura alors $u(a) = u'(a) = \cdots = u^{(n)}(a) = 0$ et d'après le résultat précédent, $\big[f(x) - P(x)\big] \underset{x \to a}{\ll} (x-a)^n$.

La formule de Taylor-Young. *Pour une fonction f ayant des dérivées jusqu'à l'ordre n, on a*

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \varphi(x)\,,$$

où $\varphi(x) \underset{x \to a}{\ll} (x-a)^n$.

Définition
Le polynôme $P(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$ s'appelle le *développement limité* ou *l'approximation à l'ordre n de f au point a*.

D'après la formule de Taylor-Young, l'approximation à l'ordre n est unique.
À l'ordre 1, le développement limité au point a est $f(a) + f'(a)(x-a)$: c'est l'approximation affine de f au point a (page 272).

2.2 Calcul des développements limités

Notation petit o : Supposons que n est un entier positif. Toute quantité qui est infiniment petite devant $(x-a)^n$ quand x tend vers a, se note $o\big[(x-a)^n\big]$, ce qui se lit « petit o de $(x-a)^n$ ».
Ainsi par exemple, on a $x^3 = o(x^2)$ (infiniments petits quand x tend vers 0) et $(x-a)^2 = o(x-a)$ (infiniments petits quand x tend vers a).
On peut calculer les développements limités au moyen de la formule de Young, mais au prix de plusieurs dérivations. Voici un premier outil de calcul bien commode.

Intégration d'un développement limité. *Supposons que le développement limité à l'ordre n de f en a est $f(x) = p_0 + p_1(x-a) + \cdots + p_n(x-a)^n + o\big[(x-a)^n\big]$. Si F est une primitive de f, alors*

$$F(x) = F(a) + p_0(x-a) + \frac{p_1}{2}(x-a)^2 + \cdots + \frac{p_n}{n+1}(x-a)^{n+1} + o\big[(x-a)^{n+1}\big]\,.$$

Démonstration. Supposons pour simplifier $a=0$, de sorte que $f(x)=p_0+p_1x+\cdots+p_nx^n+o(x^n)$.
Si F est une primitive de f, alors

$$F(x) = F(0) + \int_0^x f(t)\,dt = F(0) + p_0x + \frac{p_1}{2}x^2 + \cdots + \frac{p_n}{n+1}x^{n+1} + \int_0^x o(t^n)\,dt$$

On sait qu'en intégrant de 0 à x une fonction de t qui est négligeable devant t^n quand t tend vers 0, on obtient, quand x tend vers 0, une quantité négligeable devant $\int_0^x t^n\,dt = \frac{x^{n+1}}{n+1}$ (page 289), donc $\int_0^x o(t^n)\,dt = o(x^{n+1})$. ■

Mis à part l'intégration, on peut ajouter ou multiplier des développements limités au même point et au même ordre.

Quelques développements limités usuels au point 0

1) À l'ordre 1, on a l'approximation affine $e^x = 1 + x + o(x)$. En intégrant, on en déduit $e^x - e^0 = \int_0^x e^t\,dt = \int_0^x (1+t)\,dt + o(x^2) = x + \frac{1}{2}x^2 + o(x^2)$, c'est-à-dire

$$e^x = 1 + x + \frac{1}{2}x^2 + o(x^2)$$

En intégrant à nouveau, il vient de même $e^x - 1 = \int_0^x \left(1 + t + \frac{1}{2}t^2\right) dt + o(x^3)$, ou encore

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{2\cdot 3}x^3 + o(x^3)$$

Si l'on continue à intégrer, on obtient les développements à des ordres plus élevés.

2) Pour n entier positif, écrivons l'identité $\frac{1-x^{n+1}}{1-x} = 1 + x + \cdots + x^n$ sous la forme

$$\frac{1}{1-x} = 1 + x + \cdots + x^n + \frac{x^{n+1}}{1-x}$$

Quand x tend vers 0, $\frac{1}{1-x}$ tend vers 1, donc $\frac{x^{n+1}}{1-x} = o(x^n)$. On en déduit que le développement limité de $\frac{1}{1-x}$ à l'ordre n en 0 est :

$$\frac{1}{1-x} = 1 + x + \cdots + x^n + o(x^n)$$

En changeant x en $-x$, puis x en x^2, on trouve

$$\frac{1}{1+x} = 1 - x + x^2 - \cdots + (-1)^n x^n + o(x^n)$$
$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - \cdots + (-1)^n x^{2n} + o(x^{2n+1})$$

Dans le dernier développement, on a gagné un ordre, car on sait que tous les termes sont d'exposant pair.

3) Intégrons le développement $\frac{1}{1-x} = 1 + x + \cdots + x^{n-1} + o(x^{n-1})$.
Puisque $\int_0^x \frac{dt}{1-t} = -\ln(1-x)$, il vient

$$\ln(1-x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \cdots - \frac{1}{n}x^n + o(x^n)$$

4) On a $\frac{1}{1+x^2} = 1 - x^2 + o(x^3)$, donc $\operatorname{Arc\,tan} x = x - \frac{1}{3}x^3 + o(x^4)$.

5) On sait que $\sin x = x + o(x)$. Puisque $\int_0^x \sin t\,dt = 1 - \cos x$, on a $1 - \cos x = \frac{1}{2}x^2 + o(x^2)$, c'est-à-dire $\cos x = 1 - \frac{1}{2}x^2 + o(x^2)$. En intégrant à nouveau, on obtient $\sin x = x - \frac{1}{6}x^3 + o(x^3)$ et en poursuivant ainsi, on trouve les développements à des ordres supérieurs.

6) Cherchons le développement limité de $\sqrt{1-x}$ en 0 à l'ordre 3 : la dérivée de $\sqrt{1-x}$ est $-\frac{1}{2\sqrt{1-x}}$, elle vaut $-1/2$ en $x = 0$, donc le développement cherché s'écrit

$$\sqrt{1-x} = 1 - \frac{1}{2}x + ax^2 + bx^3 + o(x^3)$$

Élevons au carré en négligeant les termes infiniment petits devant x^3 :

$$\begin{aligned} 1-x &= 1+2\Big(-\frac{1}{2}x\Big)+\Big(\frac{1}{2}\Big)^2x^2+2ax^2+2\Big(-\frac{1}{2}x\Big)(ax^2)+2bx^3 + o(x^3) \\ &= 1 - x + \Big(2a + \frac{1}{4}\Big)x^2 + (2b - a)x^3 + o(x^3) \end{aligned}$$

Puisque le développement limité est unique, les parties polynômes sont les mêmes, donc $2a + \frac{1}{4} = 2b - a = 0$, ce qui donne $a = -\frac{1}{8}$, $b = -\frac{1}{16}$ et

$$\sqrt{1-x} = 1 - \frac{1}{2}x - \frac{1}{8}x^2 - \frac{1}{16}x^3 + o(x^3)$$

Principaux développements en 0
$\frac{1}{1-x} = 1 + x + \cdots + x^n + o(x^n)$
$e^x = 1 + x + \frac{1}{2!}x^2 + \cdots + \frac{1}{n!}x^n + o(x^n)$
$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + o(x^3)$
$\sin x = x - \frac{1}{6}x^3 + o(x^4)$, $\cos x = 1 - \frac{1}{2}x^2 + o(x^3)$, $\tan x = x + \frac{1}{3}x^3 + o(x^4)$
$\operatorname{Arc\,sin} x = x + \frac{1}{6}x^3 + o(x^4)$, $\operatorname{Arc\,tan} x = x - \frac{1}{3}x^3 + o(x^4)$
$(1-x)^{1/2} = 1 - \frac{1}{2}x - \frac{1}{8}x^2 + o(x^2)$, $(1-x)^{-1/2} = 1 + \frac{1}{2}x + \frac{3}{8}x^2 + o(x^2)$

Pratique des développements limités

Pour calculer un développement limité $p_0 + p_1(x-a) + \cdots + p_n(x-a)^n + o\big[(x-a)^n\big]$ au point a, il est commode de se ramener au point 0 en posant $X = x - a$.

Exemples

a) Cherchons le développement limité de $\frac{1}{x}$ en un point $a \neq 0$. On pose $X = x - a$ et l'on calcule le développement limité de $\frac{1}{a+X}$ au point $X = 0$. À l'ordre 2 par exemple, il vient

$$\frac{1}{a+X} = \frac{1}{a}\,\frac{1}{1+X/a} = \frac{1}{a}\left[1 - \frac{X}{a} + \frac{X^2}{a^2} + o(X^2)\right] = \frac{1}{a} - \frac{X}{a^2} + \frac{X^2}{a^3} + o(X^2)$$

$$\frac{1}{x} = \frac{1}{a} - \frac{1}{a^2}(x-a) + \frac{1}{a^3}(x-a)^2 + o\big[(x-a)^2\big]$$

b) Pour trouver le développement de $\cos x$ à l'ordre 2 au point $x = \pi/3$, posons $X = x - \pi/3$ et développons $\cos(\pi/3 + X)$ au point $X = 0$ à l'ordre 2. Puisque $\cos\pi/3 = 1/2$ et $\sin\pi/3 = \sqrt{3}/2$, il vient

$$\begin{aligned}
\cos(\pi/3 + X) &= \frac{1}{2}\cos X - \frac{\sqrt{3}}{2}\sin X \\
&= \frac{1}{2}\left[1-\frac{1}{2}X^2\right] - \frac{\sqrt{3}}{2}X + o(X^2) \\
&= \frac{1}{2} - \frac{\sqrt{3}}{2}X - \frac{1}{4}X^2 + o(X^2) \\
\cos x &= \frac{1}{2} - \frac{\sqrt{3}}{2}(x-\pi/3) - \frac{1}{4}(x-\pi/3)^2 + o\big[(x-\pi/3)^2\big]
\end{aligned}$$

Somme. On peut évidemment ajouter ou soustraire deux développements limités au même point et l'ordre du résultat est le plus petit des ordres employés.
Par exemple, au point 0, on a

$$\sin x = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 + o(x^6) \quad \text{et} \quad \tan x = x + \frac{1}{3}x^3 + o(x^3)$$

donc $\tan x - \sin x = \left(\frac{1}{3}+\frac{1}{6}\right)x^3 + o(x^3) = \frac{1}{2}x^3 + o(x^3)$. C'est le résultat qu'indique le tableau de valeurs page 265.

Produit. Pour trouver le développement à l'ordre n au point 0 d'un produit $f(x)g(x)$, on écrit les développements limités de $f(x)$ et $g(x)$ à ce même ordre n et l'on multiplie les parties polynômes en négligeant les puissances de x supérieures à n. Par exemple, on a

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + o(x^3) \quad \text{et} \quad \cos x = 1 - \frac{1}{2}x^2 + o(x^3)$$

$$\begin{aligned}
e^x \cos x &= \left[1+x+\frac{1}{2}x^2+\frac{1}{6}x^3\right]\left[1-\frac{1}{2}x^2\right] + o(x^3) \quad \text{(produit des parties polynômes)} \\
&= 1+x+\frac{1}{2}x^2+\frac{1}{6}x^3+[1+x]\left[-\frac{1}{2}x^2\right] + o(x^3) \quad \text{(suppression des ordres} > 3) \\
&= 1+x-\frac{1}{3}x^3+o(x^3)
\end{aligned}$$

Composé. Supposons que $z = g(y)$ et $y = f(x)$ ont des développements limités $z = Q(y) + o(y^n)$ et $y = P(x) + o(x^n)$ au point 0 et que l'on a $f(0) = 0$. Cette dernière condition se traduit par le fait que le terme constant du polynôme P est nul. Pour obtenir le développement de $z = g\big(f(x)\big)$ à l'ordre n au point 0, on remplace y par $P(x)$ dans le polynôme $Q(y)$ en négligeant les puissances de x supérieures à n.
Cherchons par exemple le développement à l'ordre 3 de $z = \sqrt{\cos x}$ au point 0.
On pose $y = \cos x - 1$, donc $z = \sqrt{\cos x} = \sqrt{1+y}$, et l'on a bien $y = 0$ quand $x = 0$.

$$y = -\frac{1}{2}x^2 + o(x^3) \quad \text{et} \quad z = 1 + \frac{1}{2}y - \frac{1}{8}y^2 + o(y^2)$$

Puisque $y^2 \underset{x\to 0}{\sim} \frac{1}{4}x^4$, on a $y^2 = o(x^3)$, donc aussi $o(y^2) = o(x^3)$. En remplaçant dans le développement de z, on obtient

$$\sqrt{\cos x} = z = 1 + \frac{1}{2}\left(-\frac{1}{2}x^2\right) + o(x^3) = 1 - \frac{1}{4}x^2 + o(x^3)$$

Recherche d'une partie principale et calcul de limites

Quand x tend vers une valeur a. Si l'on a un développement limité de $f(x)$ en un point a, alors $f(x)$ est équivalent au premier terme non nul de ce développement.

En effet, si le premier terme est $p_k(x-a)^k$, alors $f(x) - p_k(x-a)^k$ est une somme de fonctions négligeables devant $(x-a)^k$.

Quand x tend vers l'infini. On prend $t = \frac{1}{x}$ comme variable : alors $f(x)$ est une fonction $g(t)$ et l'on calcule le développement limité de $g(t)$ quand t tend vers 0 par valeurs positives si x tend vers $+\infty$, par valeurs négatives si x tend vers $-\infty$.

Exemples

a) Quelle est la partie principale de $f(x)=\sqrt{x^3-ax}-\sqrt{x^3-bx}$ quand x tend vers $+\infty$? Supposons $a \neq b$, sinon $f(x)=0$. Quand x tend vers $+\infty$, la partie prépondérante de $\sqrt{x^3-ax}$ est $\sqrt{x^3} = x\sqrt{x}$: on met donc $x\sqrt{x}$ en facteur, d'où

$$f(x) = x\sqrt{x}\left[\sqrt{1-a/x^2} - \sqrt{1-b/x^2}\right] = x\sqrt{x}\,u(x)$$

En posant $x = 1/t$ et en utilisant les développements limités de $\sqrt{1+t^2}$ et de $\sqrt{1-t^2}$ en 0 (tableau page 303), il vient

$$u(x) = \sqrt{1-at^2} - \sqrt{1-bt^2} = \left(1 - \frac{at^2}{2}\right) - \left(1 - \frac{bt^2}{2}\right) + o(t^2) = \frac{b-a}{2}t^2 + o(t^2) \underset{t\to 0}{\sim} \frac{b-a}{2}t^2$$

$$f(x) \underset{x\to+\infty}{\sim} x\sqrt{x}\,\frac{b-a}{2x^2} = \frac{b-a}{2\sqrt{x}}$$

On en déduit $\lim\limits_{x\to+\infty} f(x) = \lim\limits_{x\to+\infty} \frac{b-a}{2\sqrt{x}} = 0$.

b) Cherchons la partie principale de $f(x) = \frac{1}{\sin x} - \frac{1}{x}$ quand x tend vers 0. On a $f(x) = \frac{x - \sin x}{x \sin x}$ et l'on sait que $(x - \sin x) \underset{x\to 0}{\sim} \frac{x^3}{6}$. On en déduit $f(x) \underset{x\to 0}{\sim} \frac{x^3/6}{x\sin x} = \frac{x}{6}\frac{x}{\sin x} \underset{x\to 0}{\sim} \frac{x}{6}$, car $\sin x \underset{x\to 0}{\sim} x$. En conséquence, $f(x)$ tend vers 0 quand x tend vers 0.

Exemple. Pour un certain cocktail de médicaments, on sait que la quantité q de produit actif dans le tissu cible se stabilise puis décroît en fonction du temps selon une loi de diffusion de la forme $q(t) = ae^{-\alpha t} + be^{-\beta t}$, où $\beta > \alpha > 0$.
Un des problèmes en Biométrie consiste à estimer les constantes positives a, α, b et β en effectuant des mesures de $q(t)$.

Quand t devient grand, $be^{-\beta t}$ est infiniment petit devant $ae^{-\alpha t}$: en effet, $\dfrac{e^{-\beta t}}{e^{-\alpha t}}=e^{(\alpha-\beta)t}$ tend vers 0 quand t tend vers $+\infty$, car nous avons supposé $\alpha-\beta<0$. On a donc $q(t)\underset{t\to+\infty}{\sim}ae^{-\alpha t}$. Puisque $q(t)=ae^{-\alpha t}\big[1+(b/a)e^{(\alpha-\beta)t}\big]$, il vient en prenant le logarithme

$$\ln q(t)=-\alpha t+\ln a+\ln\big[1+u(t)\big],\quad \text{où } u(t)=(b/a)e^{(\alpha-\beta)t}.$$

Comme $u(t)$ tend vers 0 quand t tend vers plus l'infini, $\ln\big[1+u(t)\big]$ aussi, et $\ln q(t)\underset{t\to+\infty}{\sim}(-\alpha t+\ln a)$.

Après un certain délai qui dépend de l'ordre de grandeur (connu) des coefficients, $\ln q(t)$ se comporte comme la fonction affine $-\alpha t+\ln a$. Pour calculer une estimation de $-\alpha$ et de $\ln a$, on effectue des mesures de $q_1,q_2,\ldots$ à des instants $t_1,t_2,\ldots$ et l'on cherche la fonction affine qui approxime le mieux la relation entre les t_i et les $q_i=q(t_i)$: la solution est la droite de régression (page 211).

Après qu'on ait ainsi déterminé les nombres α et $\ln a$ (donc aussi le coefficient a), on considère

$$r(t)=q(t)-ae^{-\alpha t}=be^{-\beta t}\quad,\quad \ln r(t)=-\beta t+\ln b$$

et les mesures t_i, $\ln r_i=\ln\big(q_i-ae^{-\alpha t_i}\big)$. En calculant la droite de régression relative à ces données, on obtient les nombres $-\beta$ et $\ln b$ (voir l'exercice 4).

Application à l'étude d'une fonction au voisinage d'un point

Soit f une fonction continue ayant au point a un développement limité à l'ordre 2 :

$$f(x)=p_0+p_1(x-a)+p_2(x-a)^2+o\big[(x-a)^2\big]$$

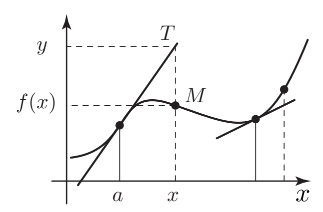
Proposition

a) *On a* $p_0=f(a)$ *et* $p_1=f'(a)$.

b) *L'équation de la tangente en* a *est* $y=p_0+p_1(x-a)$.

c) *Au voisinage de* a, *la courbe est au-dessus de sa tangente en* a *si* $p_2>0$, *en-dessous si* $p_2<0$.

Le membre de droite tend vers p_0 quand x tend vers a ; comme la fonction est continue en a, $f(x)$ tend vers $f(a)$, donc $p_0=f(a)$. On a alors $\big(f(x)-f(a)\big)\underset{x\to a}{\sim}p_1(x-a)$ et par définition, $f'(a)=p_1$. L'équation de la tangente est donc bien $y=p_0+p_1(x-a)$. Soient M et T les points d'abscisse x sur le graphe de f et sur la tangente en a. Si $p_2\neq 0$, on a $\overline{TM}=f(x)-y\underset{x\to a}{\sim}p_2(x-a)^2$. Pour x voisin de a, $\overline{TM}$ a donc le signe de $p_2(x-a)^2$, c'est-à-dire celui de p_2.



Si f est deux fois dérivable, on sait qu'en tout point a, le coefficient p_2 du développement limité est $\frac{1}{2}f''(a)$. On en déduit :

Si $f''(x)>0$ *pour tout* x, *alors en tout point, le graphe de* f *est au dessus de sa tangente.*

Si $f''(x)<0$ *pour tout* x, *alors en tout point, le graphe de* f *est en dessous de sa tangente.*

Majoration de l'erreur dans l'approximation affine

Pour une fonction deux fois dérivable, on a une égalité des accroissements finis à l'ordre 2 :

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(c), \quad \text{où } c \text{ est entre } a \text{ et } x.$$

L'erreur commise en x quand on remplace $f(x)$ par son approximation affine au point a est moindre que $M\frac{|x-a|^2}{2}$, où M est un majorant de $\left|f''(t)\right|$ quand t décrit l'intervalle d'extrémités a et x.

En effet, sur la figure précédente, on a $TM = \left|f(x) - y\right| = \frac{(x-a)^2}{2}\left|f''(c)\right|$.

3. Résolution d'équations par la méthode de Newton

Dans les applications, on rencontre le plus souvent des équations qu'on ne sait pas résoudre de façon exacte. Il faut alors disposer de méthodes permettant de trouver une bonne valeur approchée de la solution. La méthode de Newton est l'une des plus employées pour résoudre une équation $f(x)=0$ lorsque la fonction est suffisamment régulière.

Précisément, supposons que f a une dérivée seconde continue et que $f'(x) \neq 0$ sur l'intervalle où l'on cherche la solution.

Principe de la méthode. On considère la fonction $N(x) = x - \frac{f(x)}{f'(x)}$ et la suite itérative $x_{n+1} = N(x_n)$, c'est-à-dire

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{où } x_0 \text{ est une valeur initiale donnée.}$$

Un nombre s est solution de l'équation $f(x) = 0$ si et seulement si $N(s) = s$, autrement dit : les solutions de $f(x)=0$ sont les points fixes de N. Si la suite (x_n) a une limite, cette limite est un point fixe de N, donc une solution de l'équation $f(x) = 0$.

On a $N'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$, donc $N'(s) = 0$ car $f(s) = 0$.

Pour x voisin de s, $|N'(x)|$ sera voisin de 0 et en tout cas inférieur à un nombre positif $K < 1$. D'après l'inégalité des accroissements finis, on aura donc

$$|x_{n+1} - s| = \left|N(x_n) - N(s)\right| \leqslant K|x_n - s|$$

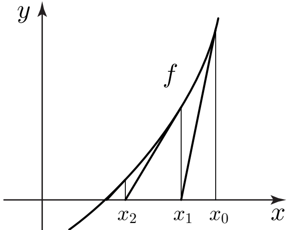
pour x_n assez proche de s. Ainsi la fonction N est contractante au voisinage du point fixe s. Le point fixe est attractif et la suite (x_n) tend vers s (page 263).

Mais puisque $N'(s) = 0$, le coefficient K est d'autant plus petit que x_n est proche de s, donc le point fixe est très attractif et la suite (x_n) converge très rapidement vers s : c'est cela qui fait l'intérêt de la méthode de Newton.

Interprétation géométrique. Au point initial x_0, la tangente au graphe de f a pour équation $y = f(x_0) + f'(x_0)(x - x_0)$. Puisque $f'(x_0) \neq 0$, cette tangente coupe l'axe des abscisses au point x tel que $0 = f(x_0) + f'(x_0)(x - x_0)$, c'est-à-dire en $x_1 = x_0 - \dfrac{f(x_0)}{f'(x_0)} = N(x_0)$.

On voit maintenant comment construire les itérés x_n :

la tangente au graphe de f au point d'abscisse x_n coupe l'axe des abscisses en x_{n+1}.



Exemple. Soit $f(x) = 4x^3 - 11x^2 + 15x - 14$. On a $f(1) = -6 < 0$ et $f(2) = 4 > 0$, donc l'équation $f(x) = 0$ possède une solution s entre 1 et 2, d'après le théorème des valeurs intermédiaires.

Prenons $x_0 = 2{,}5$ comme valeur initiale. Le tableau ci-dessous montre les premières valeurs x_n de la suite de Newton. Il est visible que cette suite a pour limite 1,75 et, dans cet exemple, on a effectivement $f(1{,}75) = 0$: le nombre $s = 1{,}75$ est solution exacte de l'équation.

n	x_n	e_n	décimales exactes
0	2,50000000000000000000000000	0,75	0
1	2,007142857142857142857142857	0,20	0
2	1,791552915934118533759009012	0,04	1
3	1,751265118850405510535778693	10^{-3}	2
4	1,750001206860097353452187869	10^{-6}	5
5	1,750000000001099252866066046	10^{-12}	11
6	1,750000000000000000000000912	10^{-24}	24

Le nombre e_n est l'erreur commise en remplaçant s par x_n, autrement dit $e_n = x_n - s$ (on ne connaît bien sûr de e_n que des valeurs approchées).

Dans le tableau, on voit qu'en passant de x_n à x_{n+1}, le nombre de décimales exactes est au moins doublé, de sorte que la convergence est très rapide. Nous allons voir que c'est une propriété générale des termes x_n de la suite de Newton, pour n assez grand.

Proposition. *Posons $e_n = x_n - s$. Quand n tend vers l'infini, $\dfrac{e_{n+1}}{e_n^2}$ tend vers $\dfrac{f''(s)}{2f'(s)}$.*

Démonstration. On a $N'(x) = \dfrac{f(x)f''(x)}{f'(x)^2}$ et puisque $f(s) = 0$, $f(x) = f'(s)(x-s) + o(x-s)$. Quand x tend vers s, $f'(x)$ tend vers $f'(s) \neq 0$, $f''(x)$ tend vers $f''(s)$, donc $N'(x) = \dfrac{f'(s)(x-s)f''(s)}{f'(s)^2} + (x-s)\,o(x-s)$. En intégrant, on obtient $N(x) - s = N(x) - N(s) =$

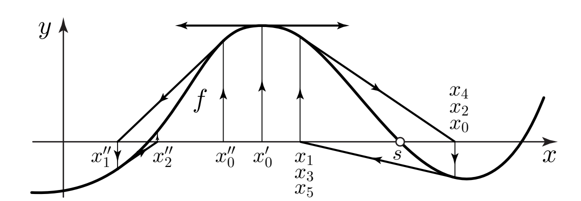
$\dfrac{1}{2}\dfrac{f''(s)}{f'(s)}(x-s)^2+o\big[(x-s)^2\big]$, d'où $\lim_{x\to s}\dfrac{N(x)-s}{(x-s)^2}=\dfrac{f''(s)}{2f'(s)}$. On a $N(x_n)=x_{n+1}$ et $\lim x_n=s$, donc $\dfrac{x_{n+1}-s}{(x_n-s)^2}$ tend vers $\dfrac{f''(s)}{2f'(s)}$. ■

Pour fixer les idées, supposons que $\left|f''(s)/2f'(s)\right|$ est de l'ordre de l'unité. À partir du rang n tel que $|e_n|$ soit de l'ordre de 10^{-1}, $|e_{n+1}|$ sera de l'ordre de $(10^{-1})^2=10^{-2}$, $|e_{n+2}|$ de l'ordre de $(10^{-2})^2=10^{-4}$, et ainsi de suite : à partir de ce rang n, le nombre de décimales exactes est en principe doublé quand on passe de x_n à x_{n+1}.

Remarque

Dans la méthode du point fixe pour une fonction contractante, nous avons montré que l'erreur diminue comme k^n, où $k<1$ est le coefficient de contraction de la fonction au voisinage du point fixe (page 263) ; si par exemple k est de l'ordre de $1/10$, on finit seulement par gagner une décimale à chaque itération. On voit que la suite de Newton converge beaucoup plus rapidement.

Précautions d'emploi. Il est important de choisir une valeur initiale x_0 suffisamment proche de la solution, faute de quoi la suite peut tendre vers l'infini, ou générer un cycle, ou devenir erratique.



On commence donc en général par estimer grossièrement la solution, afin de prendre x_0 pas trop différent de s.

Cependant, supposons par exemple que pour $x>s$, on a $f(x)>0$ et f' strictement croissante, comme sur la figure page 308 (la condition f' croissante se traduit par le fait que le graphe est au dessus de ses tangentes). Pour toute valeur initiale $x_0>s$, on a alors $s<x_1<x_0$, donc aussi $s<x_2<x_1$ et ainsi de suite : la suite (x_n) est décroissante, minorée par s, donc elle converge. Sa limite doit être une solution de l'équation $f(x)=0$, et comme on a supposé $f(x)>0$ pour tout $x>s$, la limite est s. De même, si $f(x)$ et $f''(x)$ sont négatifs pour tout $x>s$, on peut choisir $x_0>s$ quelconque. Ces conditions sont notamment satisfaites dans le cas suivant.

Si f est une fonction polynôme ayant toutes ses racines réelles et si s est la plus grande racine, la suite de Newton initialisée en $x_0>s$ a pour limite s.

4. Courbes paramétrées

Dans un plan muni d'un repère $(O;\vec{i},\vec{j})$, considérons un point M dont les coordonnées dépendent d'un paramètre t : en notant $M(t)$ la position du point pour la valeur t du paramètre et $x(t), y(t)$ ses coordonnées, on a donc

$$\overrightarrow{OM(t)} = x(t)\,\vec{i} + y(t)\vec{j}$$

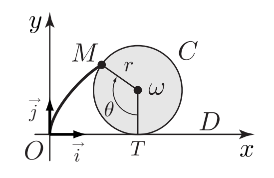
Quand t varie dans un intervalle, le point $M(t)$ décrit une courbe, appelée *courbe paramétrée*. Les fonctions $x(t)$ et $y(t)$ s'appellent les *fonctions coordonnées* de la courbe.

Exemple 1. Si les coordonnées $x(t)$ et $y(t)$ sont des fonctions affines, de la forme $x(t)=at+b$, $y(t)=ct+d$ avec $a\neq 0$ et $c\neq 0$, alors $M(t)$ décrit une droite. En effet, en éliminant t entre les égalités $x=at+b$ et $y=ct+d$, il vient $cx-ay=cb-ad$, qui est bien l'équation d'une droite D. Réciproquement, pour tout point $M\in D$ de coordonnées (x,y), on a $a(y-d)=c(x-b)$, donc en posant $t=\dfrac{x-b}{a}$, il vient $x=at+b$ et $y=ct+d$.

Exemple 2 : l'ellipse. Posons $x(t)=a\cos t$ et $y(t)=b\sin t$, où $a>0$ et $b>0$. Puisqu'on a l'identité $(\cos t)^2+(\sin t)^2=1$, l'équation de la courbe est : $(x/a)^2+(y/b)^2=1$. À nouveau, pour trouver l'équation, nous avons éliminé t entre les expressions $x(t)$ et $y(t)$. La courbe est une ellipse de centre l'origine et d'axes Ox, Oy. Quand t décrit $[0,2\pi[$, l'ellipse est parcourue une fois ; quand t décrit $\mathbb{R}$, elle est parcourue une infinité de fois. Si $a=b$, on obtient simplement un cercle de rayon a.

Exemple 3 : la cycloïde. Faisons rouler sans glissement un cercle C sur une droite D. Considérons un point M lié au cercle : au cours du mouvement, la trajectoire de M est une courbe appelée *cycloïde*.

Cherchons une équation paramétrique de la cycloïde. Pour cela, prenons comme axe des abscisses la droite D orientée dans le sens du mouvement et choisissons le repère orthonormé direct. Pour simplifier, supposons qu'à l'instant initial, le point de contact du cercle et de la droite est à l'origine. Le paramètre est l'angle θ dont a tourné le cercle. Notons r le rayon du cercle, ω son centre et T son point de contact avec D. Puisque le roulement est sans glissement, la longueur OT est égale à la longueur $r\theta$ de l'arc $\overset{\frown}{TM}$ du cercle. Le point ω décrit une droite parallèle à D à la distance r : si l'on suppose que C est dans le demi-plan supérieur, l'ordonnée de ω vaut donc toujours r. On a $\overrightarrow{OM}=\overrightarrow{OT}+\overrightarrow{T\omega}+\overrightarrow{\omega M}$, $\overrightarrow{OT}=r\theta\vec{i}$, $\overrightarrow{T\omega}=r\vec{j}$ et $\overrightarrow{\omega M}=(-r\sin\theta)\,\vec{i}+(-r\cos\theta)\vec{j}$, donc



$$\overrightarrow{OM(t)} = r(\theta-\sin\theta)\,\vec{i} + r(1-\cos\theta)\vec{j}$$

Les fonctions coordonnées sont $x(\theta)=r(\theta-\sin\theta)$, $y(\theta)=r(1-\cos\theta)$.

Comme on ne peut pas, de façon simple, tirer θ de l'une des équations et reporter dans l'autre, on étudie les propriétés de la cycloïde au moyen des deux fonctions $\theta \mapsto x(\theta)$ et $\theta \mapsto y(\theta)$.

4.1 Tangente, longueur, courbure

Tangente en un point d'une courbe paramétrée

Soit $M(t) = x(t)\vec{i} + y(t)\vec{j}$ une courbe paramétrée dont les fonctions coordonnées $x(t)$ et $y(t)$ sont dérivables. Soit t_0 une valeur du paramètre et $M(t_0)$ le point correspondant. Les approximations affines

$$x(t) = x(t_0) + x'(t_0)(t - t_0) + o(t - t_0)$$
$$y(t) = y(t_0) + y'(t_0)(t - t_0) + o(t - t_0)$$

s'écrivent vectoriellement sous la forme

$$\overrightarrow{OM(t)} = \overrightarrow{OM(t_0)} + (t - t_0)\begin{bmatrix} x'(t_0) \\ y'(t_0) \end{bmatrix} + \overrightarrow{o(t - t_0)}$$

où $\overrightarrow{o(t - t_0)}$ désigne un vecteur dont les deux coordonnées sont négligeables devant $t - t_0$ quand t tend vers t_0. On pose

$$\frac{d\overrightarrow{OM}}{dt}(t_0) = \frac{dx}{dt}(t_0)\vec{i} + \frac{dy}{dt}(t_0)\vec{j} = x'(t_0)\vec{i} + y'(t_0)\vec{j}$$

Le vecteur $\dfrac{d\overrightarrow{OM}}{dt}(t_0)$ s'appelle le *vecteur dérivé* de $\overrightarrow{OM(t)}$ en t_0

Définitions

Si le vecteur dérivé en t_0 n'est pas nul, on l'appelle *vecteur tangent* en $M(t_0)$ à la courbe. La *tangente* en $M(t_0)$ est la droite passant par $M(t_0)$ et de direction le vecteur tangent.

Si le vecteur dérivé est toujours non nul, on dit que la courbe est *régulière*.

Un point T de coordonnées (x, y) est sur la tangente si et seulement si le vecteur $\overrightarrow{M(t_0)T}$ est colinéaire au vecteur tangent. En utilisant le déterminant, cela se traduit par $\det\left(\overrightarrow{M(t_0)T}, \dfrac{d\overrightarrow{OM}}{dt}(t_0)\right) = 0$, d'où l'équation de la tangente :

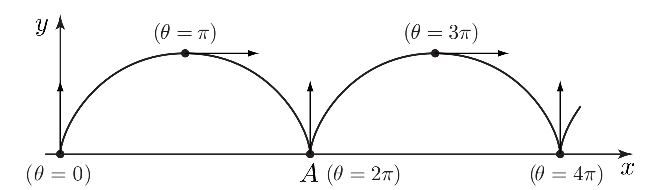
$$\begin{vmatrix} x - x(t_0) & x'(t_0) \\ y - y(t_0) & y'(t_0) \end{vmatrix} = 0$$

Si le vecteur dérivé est nul, on fait une approximation au second ordre en calculant le vecteur dérivé seconde $\dfrac{d^2\overrightarrow{OM(t)}}{dt^2} = \dfrac{d^2x}{dt^2}(t_0)\vec{i} + \dfrac{d^2y}{dt^2}(t_0)\vec{j} = x''(t_0)\vec{i} + y''(t_0)\vec{j}$. Puisqu'on a alors

$$x(t) = x(t_0) + x''(t_0)(t - t_0)^2 + o\big[(t - t_0)^2\big] \quad \text{et} \quad y(t) = y(t_0) + y''(t_0)(t - t_0)^2 + o\big[(t - t_0)^2\big],$$

la tangente en $M(t_0)$ est dirigée par ce vecteur dérivé seconde, s'il n'est pas nul. Plus généralement, le premier vecteur dérivé non nul en $M(t_0)$ donne la direction de la tangente en ce point.

Exemple. Pour la cycloïde (exemple 3 page 310), on a $\dfrac{d\overrightarrow{OM}}{d\theta} = r(1-\cos\theta)\,\vec{i} + r\sin\theta\,\vec{j}$ et ce vecteur dérivé n'est nul que si $\theta = 2k\pi$, où $k \in \mathbb{Z}$. En tout point $M(\theta)$ tel que $\theta \notin 2\pi\mathbb{Z}$, la cycloïde a une tangente. Quand $\theta = \pi$, 3π, *etc*, la tangente est horizontale, car $\sin\theta = 0$: cela correspond aux positions hautes du point M.



Pour $\theta = 0, 2\pi, 4\pi, \ldots$, le point M est sur D et le vecteur dérivé est nul. Puisque $x''(\theta) = r\sin\theta$ et $y''(\theta) = r\cos\theta$, le vecteur dérivé seconde en ces points est $r\vec{j}$, donc la tangente est verticale (points O et A sur la figure). Comme $y(\theta)$ est toujours positif ou nul, ce sont des points de rebroussement : les deux branches de la courbe y ont la même tangente.

Vecteur vitesse, vecteur accélération. Dans notre exemple, le cercle est en mouvement, donc l'angle θ est une fonction $\theta(t)$ du temps. Le vecteur dérivé $\dfrac{d\overrightarrow{OM}}{dt}$ s'appelle le *vecteur vitesse*. En notant $\dot{\theta} = \dfrac{d\theta}{dt}$, on a (dérivée d'une composée)

$$\frac{d\overrightarrow{OM}}{dt} = \frac{d\overrightarrow{OM}}{d\theta}\,\dot{\theta}$$

Ainsi le vecteur vitesse est tangent à la courbe et sa direction indique le sens de parcours. Le vecteur $\dfrac{d^2\overrightarrow{OM}}{dt^2}$, dérivé seconde par rapport au temps, s'appelle *vecteur accélération*.

Longueur d'un arc de courbe

Supposons désormais que le repère $(O;\vec{i},\vec{j})$ est orthonormé et que les coordonnées $x = x(t)$ et $y = y(t)$ sont dérivables.

A partir du point $M(t)$, donnons au paramètre un petit accroissement $\delta t > 0$ et approchons l'arc de courbe entre $M(t)$ et $M(t+\delta t)$ par le segment de droite $M(t)M(t+\delta t)$. En négligeant les quantités infiniment petites devant $(\delta t)^2$, les coordonnées de $M(t+\delta t)$ sont $x(t+\delta t) = x(t) + x'(t)\delta t$, $y(t+\delta t) = y(t) + y'(t)\delta t$ et le carré de la longueur du segment $M(t)M(t+\delta t)$ est

$M(t+\delta t)$ $y'(t)\delta t$ $M(t)$ $x'(t)\delta t$

approximation au premier ordre

$$\begin{aligned}\big[x(t+\delta t) - x(t)\big]^2 + \big[y(t+\delta t) - y(t)\big]^2 &= \big[x'(t)\delta t\big]^2 + \big[y'(t)\delta t\big]^2 \\ &= \big[x'(t)^2 + y'(t)^2\big](\delta t)^2\end{aligned}$$

En notant ds la différentielle de la longueur de l'arc, il vient

$$ds = \sqrt{x'(t)^2 + y'(t)^2}\, dt$$

Cette relation entre les différentielles ds et dt s'écrit $\frac{ds}{dt} = \sqrt{x'(t)^2 + y'(t)^2}$: elle exprime la dérivée de la longueur d'un arc.
On démontre que ce raisonnement par infiniments petits est valable pour une courbe régulière dont le vecteur tangent dépend continûment du paramètre. Dans la suite, nous supposerons toujours qu'il en est ainsi.

Définition
Choisissons un point $M_0 = M(t_0)$ de la courbe comme origine des arcs. Pour tout point $M = M(t)$, la *longueur* $s(t)$ de l'arc $\overset{\frown}{M_0M}$ a pour dérivée $\frac{ds}{dt} = \sqrt{x'(t)^2 + y'(t)^2}$. Pour tout point $M_1 = M(t_1)$, la longueur de l'arc $\overset{\frown}{M_0M_1}$ est donc $s(t_1) = \int_{t_0}^{t_1} \sqrt{x'(t)^2 + y'(t)^2}\, dt$.

La longueur de l'arc est comptée positivement si $t_1 > t_0$, négativement si $t_1 < t_0$.

Exemple : longueur d'une arche de cycloïde. Puisque nous avons choisi un repère orthonormé pour calculer les fonctions coordonnées de la cycloïde, on a

$$x'(\theta)^2 + y'(\theta)^2 = r^2(1-\cos\theta)^2 + r^2(\sin\theta)^2 = 2r^2(1-\cos\theta) = 4r^2\big[\sin(\theta/2)\big]^2$$

$$\frac{ds}{d\theta} = \sqrt{x'(\theta)^2 + y'(\theta)^2} = 2r\sin(\theta/2)\ , \quad \text{pour } 0 \leqslant \theta \leqslant 2\pi.$$

Prenons le point O ($t_0 = 0$) comme origine des arcs. Si $M(\theta)$ est un point de l'arche $\overset{\frown}{OA}$ (figure page 312), alors θ est entre 0 et 2π et la longueur de l'arc $\overset{\frown}{OM(\theta)}$ est

$$s(\theta) = 2r\int_0^\theta \sin(u/2)\, du = 2r\Big[-2\cos(u/2)\Big]_0^\theta = 4r\big[1-\cos(\theta/2)\big]$$

Pour $\theta = 2\pi$, on trouve que la longueur de l'arche $\overset{\frown}{OA}$ est $8r$.

L'abscisse curviligne

Il est naturel de paramétrer les points M de la courbe par la longueur s de l'arc $\overset{\frown}{M_0M}$. Cela est possible, car la fonction longueur est strictement croissante (nous avons supposé que $x'(t)$ et $y'(t)$ ne sont jamais nuls en même temps, donc $\frac{ds}{dt}$ est strictement positif). Ainsi s et t sont fonctions l'un de l'autre et les coordonnées x et y des points de la courbe sont aussi des fonctions de s. On dit que le paramètre s est *l'abscisse curviligne* de la courbe.
On dérive les coordonnées par rapport à l'abscisse curviligne au moyen des formules

$$\frac{dx}{ds} = \frac{dx}{dt}\frac{dt}{ds} \quad \text{et} \quad \frac{dy}{ds} = \frac{dy}{dt}\frac{dt}{ds}, \quad \text{où } \frac{dt}{ds} = \left(\frac{ds}{dt}\right)^{-1} = \big[x'(t)^2 + y'(t)^2\big]^{-1/2}$$

Il vient alors

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = \left[x'(t)^2 + y'(t)^2\right] \left(\frac{dt}{ds}\right)^2 = 1$$

Avec l'abscisse curviligne comme paramètre, le vecteur tangent est

$$\frac{d\overrightarrow{OM}}{ds} = \frac{dx}{ds}\vec{i} + \frac{dy}{ds}\vec{j}$$

Puisqu'on a supposé que le repère est orthonormé, le carré de la norme euclidienne est la somme des carrés des coordonnées ; d'après la formule précédente, on a donc

$$\left\| \frac{d\overrightarrow{OM}}{ds} \right\| = 1$$

Le vecteur $\dfrac{d\overrightarrow{OM}}{ds}$ s'appelle le *vecteur tangent unitaire.*

Courbure

On mesure la courbure d'un arc de courbe $\overparen{MM'}$ en comparant l'angle α fait par les tangentes en M et M' avec la longueur L de l'arc : la courbure moyenne de l'arc est par définition le rapport α/L.

D'après nos hypothèses, la courbe possède en tout point $M(t)$ un vecteur tangent $\dfrac{d\overrightarrow{OM}}{dt}$. Notons φ l'angle entre l'axe Ox et la tangente en M, plus précisément l'écart angulaire des vecteurs $\vec{i}, \dfrac{d\overrightarrow{OM}}{dt}(t)$ (définition page 225).

Si $M' = M(t')$ est un autre point de la courbe, l'angle des vecteurs $\vec{i}, \dfrac{d\overrightarrow{OM}}{dt}(t')$ est φ' et l'angle des tangentes est $\varphi' - \varphi$.

Choisissons un point M_0 de la courbe comme origine des arcs : les arcs $\overparen{M_0M}$ et $\overparen{M_0M'}$ ont des longueurs s et s', la longueur de l'arc $\overparen{MM'}$ est $s' - s$ et sa courbure moyenne est $\dfrac{\varphi' - \varphi}{s' - s}$. Quand M' tend vers M, la courbure moyenne a donc pour limite $\dfrac{d\varphi}{ds}$, si cette limite existe.

Supposons que les fonctions $x(t)$ et $y(t)$ sont deux fois dérivables.

Définitions

La *courbure* au point M est $K = \dfrac{d\varphi}{ds}$, où φ est l'écart angulaire des vecteurs $\vec{i}, \dfrac{d\overrightarrow{OM}}{ds}$. Si la courbure en M n'est pas nulle, le nombre $1/|K|$ s'appelle le *rayon de courbure* en M.

Intuitivement, une courbure nulle en un point correspond à un « rayon de courbure infini », c'est-à-dire à un arc « plat » en ce point.

Exemple. Dans le cas d'un cercle de rayon R centré à l'origine (exemple 2 page 310), la tangente au point M de coordonnées $(x, y) = (R\cos t, R\sin t)$ fait l'angle $\varphi = t + \pi/2$ avec l'axe Ox. En choisissant $M(0)$ comme origine des arcs, on a

$s(t) = Rt$, donc la courbure en M est $\dfrac{d\varphi}{ds} = \dfrac{d(t+\pi/2)}{Rdt} = \dfrac{1}{R}$. Le rayon de courbure en un point d'un cercle est donc le rayon du cercle.

Calcul de la courbure. Notons $\vec{T} = \dfrac{d\overrightarrow{OM}}{ds}$ le vecteur tangent unitaire au point M. Puisque le repère est orthonormé, il s'écrit $\vec{T} = \cos\varphi\,\vec{i} + \sin\varphi\vec{j}$, où φ est l'angle $\widehat{\vec{i},\vec{T}}$. On a donc

$$\frac{dx}{ds} = \cos\varphi \quad \text{et} \quad \frac{dy}{ds} = \sin\varphi$$

Le vecteur $\vec{N} = \cos(\varphi+\pi/2)\,\vec{i} + \sin(\varphi+\pi/2)\vec{j} = -\sin\varphi\,\vec{i} + \cos\varphi\vec{j}$, orthogonal à $\vec{T}$, s'appelle le *vecteur normal* en M.

- Le vecteur normal $\vec{N}$ est de norme 1 et la base $\left(\vec{T},\vec{N}\right)$ est orthonormée directe.
- La droite passant par M et orientée selon $\vec{N}$ s'appelle la *normale principale* ; elle est orthogonale à la courbe en M.

Dérivons le vecteur $\vec{T}$ par rapport à s. Il vient

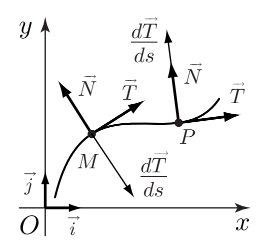
$$\frac{d\vec{T}}{ds} = (-\sin\varphi)\frac{d\varphi}{ds}\vec{i} + (\cos\varphi)\frac{d\varphi}{ds}\vec{j} = \frac{d\varphi}{ds}\vec{N}$$

Par définition de la courbure, on en déduit :

la courbure en M est le nombre K tel que $\dfrac{d\vec{T}}{ds} = K\vec{N}$

Ainsi, la courbure en M est la mesure algébrique du vecteur $\dfrac{d\vec{T}}{ds}$ sur la normale principale.

Le vecteur $\dfrac{d\vec{T}}{ds}$ est toujours dirigé vers l'intérieur de la courbe. Si la courbure en M est positive, le vecteur normal pointe lui-aussi vers l'intérieur de la courbe ; si la courbure est négative, le vecteur normal pointe vers l'extérieur.
Sur la figure, la courbure est négative en M et positive en P.



Exemple. Dans le cas de la cycloïde, on a $\dfrac{d\overrightarrow{OM}}{d\theta} = r(1-\cos\theta)\vec{i} + r\sin\theta\vec{j}$ et en supposant $0 < \theta < 2\pi$, on sait que $ds = 2r\sin(\theta/2)\,d\theta$. Calculons le vecteur tangent unitaire.

$$\vec{T} = \frac{d\overrightarrow{OM}}{ds} = \frac{d\overrightarrow{OM}}{d\theta}\frac{d\theta}{ds} = \frac{r(1-\cos\theta)}{2r\sin(\theta/2)}\vec{i} + \frac{r\sin\theta}{2r\sin(\theta/2)}\vec{j} = \sin(\theta/2)\,\vec{i} + \cos(\theta/2)\vec{j}$$

Le vecteur normal est $\vec{N} = -\cos(\theta/2)\,\vec{i} + \sin(\theta/2)\vec{j}$ et

$$\frac{d\vec{T}}{ds} = \frac{d\vec{T}}{d\theta}\frac{d\theta}{ds} = \left[\frac{1}{2}\cos(\theta/2)\,\vec{i} - \frac{1}{2}\sin(\theta/2)\vec{j}\right]\frac{1}{2r\sin(\theta/2)} = -\frac{1}{4r\sin(\theta/2)}\vec{N}$$

Au point $M(\theta)$, la cycloïde a pour courbure $K = -\dfrac{1}{4r\sin(\theta/2)}$. La courbure est minimum au sommet de l'arche ($\theta = \pi$). Remarquons qu'en un point d'ordonnée quelconque $y = r(1-\cos\theta) > 0$, la courbure est $K = -1/\sqrt{8ry}$.

Cas d'une courbe $y = f(x)$

Chaque point du graphe de f a pour coordonnées $\big(x, f(x)\big)$, donc est repéré par son abscisse : le graphe d'une fonction de x est une courbe paramétrée par x.

- Si f a une dérivée continue, le vecteur tangent au point $M(x)$ a pour coordonnées $\frac{dx}{dx} = 1$, $\frac{dy}{dx} = f'(x)$ et l'on a $\frac{ds}{dx} = \sqrt{1 + f'(x)^2}$.
- Si $f''(x) \neq 0$, le rayon de courbure en $M(x)$ est $\dfrac{\big[1 + f'(x)^2\big]^{3/2}}{\big|f''(x)\big|}$.
- Si $f''(x) = 0$, la courbure en $M(x)$ est nulle.

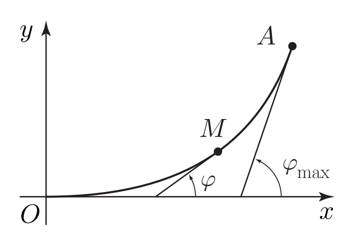
On voit que la courbure ne dépend que des deux premières dérivées de f.

4.2 Application aux bretelles de raccordement

Il s'agit de trouver la forme d'une voie routière de raccordement entre une ligne droite (courbure nulle) et un arc ayant une courbure non nulle. Ce problème se pose pour le tracé d'une ligne ferroviaire ou d'une sortie d'autoroute : si en sortant d'une ligne droite, on emprunte immédiatement une trajectoire de rayon de courbure R, l'accélération normale d'un véhicule roulant à la vitesse v passe brutalement de 0 à v^2/R, ce que l'on cherche à éviter.

Dans le cadre autoroutier, on suppose habituellement que pendant le changement de direction, le conducteur roule à une vitesse constante v et imprime au volant une rotation régulière, de sorte que l'angle de braquage des roues est proportionnel à la durée du mouvement.

- Entre des instants voisins t et $t + dt$, le véhicule a tourné d'un angle $d\varphi = \alpha dt$, où α est proportionnel à l'angle de braquage.
- En prenant comme instant initial $t = 0$ celui où le véhicule quitte la ligne droite (point O), on a alors $\alpha = at$, où a est une constante positive, donc $\frac{d\varphi}{dt} = at$.
- Notons s l'abscisse curviligne de la courbe comptée à partir du point O ($s = 0$ si $t = 0$). La vitesse du véhicule à l'instant t est $\frac{ds}{dt} = v$, donc $s = vt$.



En tout point de la courbe, la courbure est $K = \frac{d\varphi}{ds} = \frac{d\varphi}{dt}\frac{dt}{ds} = \frac{at}{v} = \frac{a}{v^2}s$; cette valeur est nulle à l'instant initial et varie proportionnellement à la distance parcourue. En notant $R = 1/K$ le rayon de courbure, il vient

$$Rs = v^2/a$$

Cette relation caractéristique de la courbe exprime que le rayon de courbure est inversement proportionnel à la distance parcourue.

Prenons un repère orthonormé d'origine O, l'axe Ox étant la ligne droite et l'axe Oy étant dirigé vers la sortie. L'angle φ entre Ox et le vecteur tangent à la courbe est l'angle dont a tourné le véhicule. Puisque $\frac{d\varphi}{ds} = \frac{a}{v^2}s$, il vient $\varphi = \frac{a}{2v^2}s^2$, d'où

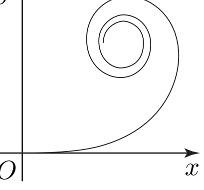
$$\frac{dx}{ds} = \cos\varphi = \cos\big[(a/2v^2)s^2\big] \quad \text{et} \quad \frac{dy}{ds} = \sin\varphi = \sin\big[(a/2v^2)s^2\big]$$

En intégrant, on trouve les fonctions coordonnées $x(s)$ et $y(s)$ de la courbe. Puisque $x = y = 0$ quand $s = 0$, on obtient ainsi la paramétrisation

$$x(s) = \int_0^s \cos\big((a/2v^2)u^2\big)\,du \quad \text{et} \quad y(s) = \int_0^s \sin\big((a/2v^2)u^2\big)\,du$$

En général, on impose l'angle $\varphi_{\max}$ dont la courbe permet de tourner, ou bien son rayon de courbure $R_{\max}$ à l'extrémité. La longueur L de l'arc est alors donnée par les relations $\varphi_{\max} = (a/2v^2)L^2$ ou $R_{\max}L = v^2/a$. L'extrémité A de la bretelle de raccordement a donc une position déterminée, de coordonnées de $x(L), y(L)$.

Ces intégrales (dites de Fresnel) ne s'expriment pas au moyen des fonctions usuelles ; pour dessiner la courbe, appelée spirale de Cornu, il faut faire un calcul numérique approché de $x(s)$ et $y(s)$ (voir le chapitre 11).



une spirale de Cornu

5. Calcul de primitives

Il est utile de savoir calculer quelques types courants d'intégrales. Rappelons les règles essentielles :

- $\int_a^b f(t)\,dt = F(b) - F(a)$ si F est une primitive de f.
- Si α et β sont des constantes, $\int \big[\alpha f(t) + \beta g(t)\big]\,dt = \alpha \int f(t)\,dt + \beta \int g(t)\,dt$.

5.1 Méthodes générales

On dispose de quatre méthodes générales pour calculer une intégrale $I = \int f(t)\,dt$.

La fonction à intégrer est une dérivée connue

Exemples

1) $\int (at+b)^\alpha\,dt = \frac{1}{a(\alpha+1)}(ax+b)^{\alpha+1}$ si $\alpha \neq -1$ et $a \neq 0$.

La dérivée de $(ax+b)^{\alpha+1}$ est en effet $a(\alpha+1)(ax+b)^\alpha$.

2) $\int \tan t\,dt = \int \frac{\sin t}{\cos t}\,dt = \int \frac{-\cos'(t)}{\cos t}\,dt = -\ln\big|\cos x\big|$. On en déduit par exemple $\int_0^{\pi/4} \tan t\,dt = \big[-\ln\cos x\big]_0^{\pi/4} = (\ln 2)/2$.

3) On a $\int (\tan t)^2\,dt = \tan x - x$, car la dérivée de $\tan x$ est $1 + (\tan x)^2$.

4) $I = \int \frac{t}{\sqrt{a+t^2}}\,dt = \int \frac{2t}{2\sqrt{a+t^2}}\,dt = \int \frac{u'(t)}{2\sqrt{u(t)}}\,dt$, où $u(x) = a + x^2$. La fonction sous le signe intégrale est la dérivée de $\sqrt{u(x)}$, donc $I = \sqrt{a+x^2}$.

Décomposer en somme

C'est ainsi que l'on peut calculer les intégrales $\int (\cos t)^n \, dt$ et $\int (\sin t)^n \, dt$ pour n entier positif. Par exemple, grâce à la formule $\cos 2x = 2(\cos x)^2 - 1 = 1 - 2(\sin x)^2$, il vient

$$2\int (\cos t)^2 \, dt = \int (1 + \cos 2t) \, dt = x + \int \cos 2t \, dt = x + \frac{1}{2}\sin 2x$$

Nous allons voir que $(\cos x)^n$ est une somme de fonctions $\cos kx$, où les nombres k qui interviennent sont des entiers positifs ou nuls au plus égaux à n. De même, $(\sin x)^n$ est une somme de fonctions $\sin kx$.

Faisons un calcul dans les nombres complexes en introduisant $z = e^{\boldsymbol{i}\,x} = \cos x + \boldsymbol{i} \sin x$. Puisque $z^{-1} = e^{-\boldsymbol{i}\,x} = \cos x - \boldsymbol{i} \sin x$, on a

$$2\cos x = z + z^{-1} \quad \text{et} \quad 2\,\boldsymbol{i}\sin x = z - z^{-1}.$$

En élevant par exemple à la puissance 4 par la formule du binôme (page 61), on obtient

$$\begin{aligned} 2^4(\cos x)^4 &= (z + z^{-1})^4 = z^4 + \binom{4}{1} z^3 z^{-1} + \binom{4}{2} z^2 z^{-2} + \binom{4}{3} z z^{-3} + z^{-4} \\ &= z^4 + 4z^2 + 6 + 4z^{-2} + z^{-4} = (z^4 + z^{-4}) + 4(z^2 + z^{-2}) + 6 \end{aligned}$$

D'après la formule de Moivre (page 37), $z^2 = \cos 2x + \boldsymbol{i}\sin 2x$ et $z^{-2} = \cos 2x - \boldsymbol{i}\sin 2x$, donc $z^2 + z^{-2} = 2\cos 2x$. De même, $z^4 + z^{-4} = 2\cos 4x$, donc il vient $2^4(\cos x)^4 = 2\cos 4x + 8\cos 2x + 6$, ou encore

(1) $$(\cos x)^4 = \frac{1}{8}\cos 4x + \frac{1}{2}\cos 2x + \frac{3}{8}$$

En intégrant, on obtient

$$\int (\cos t)^4 \, dt = \frac{1}{8}\int \cos 4t \, dt + \frac{1}{2}\int \cos 2t \, dt + \frac{3}{8}x = \frac{\sin 4x}{32} + \frac{\sin 2x}{4} + \frac{3x}{8}$$

Pour l'intégrale de $(\sin x)^3$, écrivons $(z - z^{-1})^3 = (2\,\boldsymbol{i}\sin x)^3 = -2^3\,\boldsymbol{i}(\sin x)^3$, puis

$$\begin{aligned} -8\,\boldsymbol{i}(\sin x)^3 &= z^3 - 3z^2 z^{-1} + 3z z^{-2} - z^{-3} = z^3 - 3z + 3z^{-1} - z^{-3} \\ &= (z^3 - z^{-3}) - 3(z - z^{-1}) \end{aligned}$$

Puisque $z^3 = \cos 3x + \boldsymbol{i}\sin 3x$ et $z^{-3} = \cos 3x - \boldsymbol{i}\sin 3x$, on a $-8\boldsymbol{i}(\sin x)^3 = 2\boldsymbol{i}\sin 3x - 3(2\boldsymbol{i}\sin x)$, d'où

(2) $$(\sin x)^3 = -\frac{\sin 3x}{4} + \frac{3\sin x}{4},$$

$$\int (\sin t)^3 \, dt = \frac{\cos 3x}{12} - \frac{3\cos x}{4} \quad \text{et} \quad \int_0^{\pi/2} (\sin t)^3 \, dt = \left[\frac{\cos 3x}{12} - \frac{3\cos x}{4}\right]_0^{\pi/2} = 2/3.$$

Des identités comme (1) ou (2) s'appellent des *formules de linéarisation*. Elles permettent d'exprimer $(\sin x)^n (\cos x)^m$ comme une somme de produits $\sin px \cos qx$. Puisqu'on a

$$2\sin px \cos qx = \sin(p+q)x + \sin(p-q)x\,,$$

on peut ainsi en principe calculer une primitive d'un produit $(\sin x)^n (\cos x)^m$.

Intégration par parties

La formule d'intégration par parties s'écrit $\int u(t)v'(t)\,dt = u(x)v(x) - \int u'(t)v(t)\,dt$.

Elle est valable pour des fonctions u et v à dérivée continue et signifie simplement que la dérivée du produit uv est $u'v + uv'$. Avec des bornes dans les intégrales, la formule devient

$$\int_a^b u(t)v'(t)\,dt = \big[u(x)v(x)\big]_a^b - \int_a^b u'(t)v(t)\,dt\,.$$

Par intégration par parties, on peut calculer des intégrales comme $\int t^2 e^{at}\,dt$, $\int t^3 \ln t\,dt$, $\int \arcsin t\,dt$ ou $\int \arctan t\,dt$.

Exemples

1) Dans l'intégrale $I_n = \int t^n e^{at}\,dt$ (où $a \neq 0$), posons $u = t^n$ et $v' = e^{at}$. Il vient $u' = nt^{n-1}$, $v = \dfrac{1}{a}e^{at}$, donc

$$I_n = x^n \frac{1}{a} e^{ax} - \int nt^{n-1} \frac{1}{a} e^{at}\,dt = \frac{x^n e^{ax}}{a} - \frac{n}{a} I_{n-1}$$

Si n est un entier positif, le calcul de I_n se ramène de proche en proche à celui de $I_0 = \int e^{at}\,dt = \dfrac{e^{ax}}{a}$. On a ainsi $I_1 = \dfrac{ax-1}{a^2}e^{ax}$ et $I_2 = \dfrac{a^2x^2 - 2ax + 2}{a^3}e^{ax}$.

2) Pour $I = \int t^3 \ln t\,dt$, on prend $u = \ln t$ et $v' = t^3$, donc $u' = 1/t$ et $v = t^4/4$; il vient $I = (\ln x)\dfrac{x^4}{4} - \int \dfrac{1}{t}\dfrac{t^4}{4}\,dt = \dfrac{x^4 \ln x}{4} - \dfrac{1}{4}\int t^3\,dt = \dfrac{x^4 \ln x}{4} - \dfrac{x^4}{16}$.

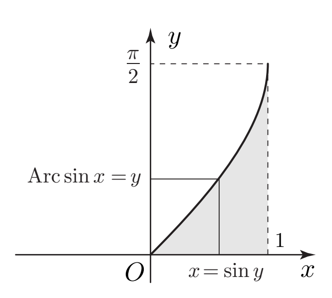
3) Calculons le nombre $\int_0^1 \operatorname{Arc\,sin} t\,dt$ en prenant $u = \operatorname{Arc\,sin} t$ et $v' = 1$, donc $v = t$ et $u' = \dfrac{1}{\sqrt{1-t^2}}$. On a $\int \operatorname{Arc\,sin} t\,dt = x \operatorname{Arc\,sin} x - \int \dfrac{t}{\sqrt{1-t^2}}\,dt$.
Dans cette dernière intégrale, le numérateur est, à un facteur $-1/2$ près, la dérivée de $w = 1 - t^2$:

$$\frac{t}{\sqrt{1-t^2}} = \frac{2t}{2\sqrt{1-t^2}} = -\frac{w'}{2\sqrt{w}} = -(\sqrt{w})'$$

On en déduit $\int \operatorname{Arc\,sin} t\,dt = x \operatorname{Arc\,sin} x + \sqrt{1-x^2}$ et

$$\int_0^1 \operatorname{Arc\,sin} t\,dt = \big[x \operatorname{Arc\,sin} x\big]_0^1 + \Big[\sqrt{1-x^2}\Big]_0^1$$
$$= \Big(\frac{\pi}{2} - 0\Big) + (0-1) = \frac{\pi}{2} - 1\,.$$

Le dessin ci-contre montre le graphe de Arc sinus et nous venons de calculer l'aire grisée. L'aire comprise entre le graphe et l'axe des ordonnées est $\int_0^{\pi/2} \sin t\,dt = 1$ et l'aire totale du rectangle est bien $(\pi/2 - 1) + 1 = \pi/2$.



Changement de variable

Supposons que $f(t)$ s'écrit $f(t) = g\big[u(t)\big]u'(t)$, où g est une certaine fonction. Alors si G est une primitive de g, la dérivée de $t \mapsto G\big[u(t)\big]$ est $g\big[u(t)\big]u'(t) = f(t)$, donc

$$\int f(t)\,dt = \int g\big[u(t)\big]u'(t)\,dt = G\big[u(x)\big]$$

Pratiquement : pour faire le changement de variable $u(t)$ dans l'intégrale,

- on écrit $f(t)$ sous la forme $g(u)$,
- on calcule $du = u'(t)\,dt$ et l'on a alors $\int f(t)\,dt = \int g(u)\,du$.

Exemples

1) Calculons ainsi l'intégrale $\int (\cos t)^n (\sin t)^m\,dt$ lorsque les exposants sont entiers positifs et que l'un d'eux, n par exemple, est impair. Posons $n = 2p+1$ et faisons le changement de variable $u = \sin t$. On a $du = \cos t\,dt$ et

$$(\cos t)^{2p+1}(\sin t)^m\,dt = (\cos t)^{2p}(\sin t)^m(\cos t\,dt) = (1-u^2)^p\,u^m\,du\,.$$

Intégrons en indiquant en haut du signe intégrale le nom de la variable à utiliser dans la primitive :

$$\int^x (\cos t)^n(\sin t)^m\,dt = \int^{\sin x} (1-u^2)^p\,u^m\,du$$

La dernière intégrale est facile à calculer car il s'agit d'une primitive de fonction polynôme. Il ne faut pas oublier d'y remplacer finalement u par $\sin x$.
Par exemple, on a

$$\begin{aligned}\int^x (\cos t)^3(\sin t)^4\,dt &= \int^x (\cos t)^2(\sin t)^4\;\cos t dt\\ &= \int^{\sin x}(1-u^2)\,u^4\,du = \int (u^4-u^6)\,du\\ &= \frac{u^5}{5}-\frac{u^7}{7} = \frac{(\sin x)^5}{5}-\frac{(\sin x)^7}{7}\,.\end{aligned}$$

2) ► Calculons $\int te^{-t^2}\,dt$ par le changement de variable $u = t^2$. Puisque $du = 2t\,dt$, on a $\int^x e^{-t^2}(t\,dt) = \frac{1}{2}\int^{x^2} e^{-u}\,du$, donc

$$\int^x te^{-t^2}\,dt = -\frac{1}{2}e^{-x^2} \quad \text{et} \quad \int_a^b te^{-t^2}\,dt = \frac{e^{-a^2}-e^{-b^2}}{2}$$

► De même, $\int^x t^3e^{-t^2}\,dt = \frac{1}{2}\int^x t^2e^{-t^2}(2t\,dt) = \frac{1}{2}\int^{x^2} ue^{-u}\,du$, et en intégrant par parties :

$$\int^x t^3e^{-t^2}\,dt = \frac{1}{2}u(-e^{-u}) - \frac{1}{2}\int -e^{-u}\,du = -\frac{x^2+1}{2}e^{-x^2}\,.$$

5.2 Intégrales de la forme $\int e^{at}\cos bt\,dt$ ou $\int e^{at}\sin bt\,dt$

Introduisons les nombres complexes

$$e^{at}\cos bt + \boldsymbol{i}\,e^{at}\sin bt = e^{at}(\cos bt + \boldsymbol{i}\sin bt) = e^{at}e^{\boldsymbol{i}\,bt} = e^{(a+b\,\boldsymbol{i})t}$$

et étendons le calcul des dérivées et des intégrales aux fonctions à valeurs complexes. Une fonction à valeurs complexes est de la forme $F(x) = U(x) + \boldsymbol{i}\,V(x)$, où $U(x)$ et $V(x)$ sont des fonctions à valeurs réelles. Définissons la dérivée de F en posant

$$F'(x) = U'(x) + \boldsymbol{i}\,V'(x)\ , \quad \text{si } U \text{ et } V \text{ sont dérivables.}$$

Proposition. *Si λ est un nombre complexe, la dérivée de la fonction $x \mapsto e^{\lambda x}$ est $\lambda e^{\lambda x}$.*

Démonstration. Posons $\lambda = a + b\,\boldsymbol{i}$, où a et b sont réels, et $f(x) = e^{\lambda x} = e^{(a+\boldsymbol{i}\,b)x} = e^{ax}\cos bx + \boldsymbol{i}\,e^{ax}\sin bx$. En dérivant, il vient

$$\begin{aligned} f'(x) &= ae^{ax}\cos bx - be^{ax}\sin bx + \boldsymbol{i}(ae^{ax}\sin bx + be^{ax}\cos bx)\\ &= e^{ax}\big[(a\cos bx - b\sin bx) + \boldsymbol{i}(a\sin bx + b\cos bx)\big]\\ &= e^{ax}(a+b\,\boldsymbol{i})(\cos bx + \boldsymbol{i}\sin bx) = (a+b\,\boldsymbol{i})e^{ax}e^{\boldsymbol{i}\,bx} = (a+b\,\boldsymbol{i})e^{(a+\boldsymbol{i}\,b)x} = \lambda e^{\lambda x}\,. \end{aligned}$$

■

Une exponentielle se dérive donc de la même manière, que l'exposant soit réel ou complexe. On en déduit que si $\lambda \neq 0$, la fonction $x \mapsto \frac{1}{\lambda}e^{\lambda x}$ est une primitive de $e^{\lambda x}$:

$$\int e^{(a+b\,\boldsymbol{i})t}\,dt = \frac{1}{a+b\,\boldsymbol{i}}e^{(a+b\,\boldsymbol{i})x}$$

En posant $u(x) = e^{ax}\cos bx$ et $v(x) = e^{ax}\sin bx$, il vient donc

$$\begin{aligned} \int u(t)\,dt + \boldsymbol{i}\int v(t)\,dt &= \frac{1}{a+b\,\boldsymbol{i}}e^{(a+b\,\boldsymbol{i})x} = \frac{a-b\,\boldsymbol{i}}{a^2+b^2}e^{ax}(\cos bx + \boldsymbol{i}\sin bx)\\ &= \frac{e^{ax}}{a^2+b^2}\Big[(a\cos bx + b\sin bx) + (-b\cos bx + a\sin bx)\,\boldsymbol{i}\Big] \end{aligned}$$

$$\int e^{at}\cos bt\,dt = \frac{e^{ax}(a\cos bx + b\sin bx)}{a^2+b^2}\quad ,\quad \int e^{at}\sin bt\,dt = \frac{e^{ax}(-b\cos bx + a\sin bx)}{a^2+b^2}$$

5.3 Intégrale d'une fonction rationnelle

Une fonction rationnelle est de la forme $\frac{P(x)}{Q(x)}$, où P et Q sont des fonctions polynômes. Contentons-nous de traiter les cas les plus couramment rencontrés.

1. $Q(x) = (x-a)^n$. Dans l'intégrale $\int \frac{P(t)}{(t-a)^n}\,dt$, on fait le changement de variable $u = t - a$, donc $du = dt$ et $\frac{P(t)}{(t-a)^n}\,dt = \frac{P(u+a)}{u^n}\,du$. Puisque P est un polynôme, $\frac{P(u+a)}{u^n}$ est une somme de puissances positives ou négatives de u.

Exemple. Par le changement de variable $u = t + 1$, on a

$$\int \frac{t^3}{(t+1)^2}\,dt = \int^{x+1} \frac{(u-1)^3}{u^2}\,du = \int (u - 3 + 3u^{-1} - u^{-2})\,du$$
$$= \frac{u^2}{2} - 3u + 3\ln|u| + \frac{1}{u} = \frac{(x+1)^2}{2} - 3(x+1) + 3\ln|x+1| + \frac{1}{x+1}$$

2. $\boldsymbol{\dfrac{P}{Q} = \dfrac{x+p}{x^2+bx+c}}$ **.** En écrivant que $x^2 + bx$ est le début d'un carré, on a

$$x^2 + bx + c = (x + b/2)^2 + d \ , \quad \text{où } d = c - b^4/4 \text{ est supposé non nul.}$$

Pour calculer l'intégrale, faisons le changement de variable $u = t + b/2$. Il vient $dt = du$, $t + p = u + p - b/2$ et

$$\int \frac{t+p}{t^2+bt+c}\,dt = \int^{x+b/2} \frac{u+p-b/2}{u^2+d}\,du = \int \frac{u}{u^2+d}\,du + (p - b/2)\int \frac{du}{u^2+d}$$

► Dans la première intégrale, le numérateur est, à un facteur $1/2$ près, la dérivée du dénominateur : cette intégrale est donc $\frac{1}{2}\ln|u^2+d| = \frac{1}{2}\ln|x^2+bx+c|$.

► Pour calculer la seconde intégrale, on pose $d = r^2$ si $d > 0$ et $d = -r^2$ sinon, et l'on obtient une intégrale usuelle (page 298).

Exemple. Calculons $I = \int \frac{2t+1}{t^2-2t+3}\,dt$. Puisque $x^2 - 2x + 3 = (x-1)^2 + 2$, on pose $u = t - 1$. On a alors $2t + 1 = 2u + 3$ et il vient

$$I = \int^{x-1} \frac{2u+3}{u^2+2}\,du = \int \frac{2u\,du}{u^2+2} + 3\int \frac{du}{u^2+2}$$
$$= \ln(u^2+2) + 3\frac{1}{\sqrt{2}}\,\text{Arc tan}\,\frac{u}{\sqrt{2}} = \ln(x^2 - 2x + 3) + \frac{3}{\sqrt{2}}\,\text{Arc tan}\,\frac{x-1}{\sqrt{2}}$$

3. Réduction du degré du numérateur. Supposons que le degré de P est supérieur à celui de Q. Faisons la division euclidienne de P par Q (page 49) en appelant E le polynôme quotient et R le reste : on a $P = EQ + R$, où R est de degré inférieur à celui de Q, d'où $\frac{P}{Q} = E + \frac{R}{Q}$.

Puisqu'il est très simple d'intégrer le polynôme $E(x)$, on est ramené à calculer une primitive de la fraction $\frac{R(x)}{Q(x)}$ où le numérateur est de degré inférieur à celui du dénominateur.

Exemple. Posons $f(x) = \frac{x^3}{x^2+1}$. La division de x^3 par x^2+1 s'écrit $x^3 = x(x^2+1) - x$, donc $f(x) = x - \frac{x}{x^2+1}$ et $\int \frac{t^3}{t^2+1}\,dt = \frac{x^2}{2} - \frac{1}{2}\ln(x^2+1)$.

4. Décomposition en somme. Supposons que $Q = AB$ est le produit de deux polynômes n'ayant aucune racine commune (réelle ou complexe). On peut alors trouver des polynômes U et V tels que $AU + BV = 1$, $\deg U < \deg B$ et $\deg V < \deg A$ (proposition page 50). Il vient alors $\dfrac{P}{Q} = \dfrac{PAU + PBV}{AB} = \dfrac{PU}{B} + \dfrac{PV}{A}$.
Exploitons cela lorsque le dénominateur Q a une factorisation très simple.

A) $\boldsymbol{Q(x) = (x-a)(x-b)}$ **, où** $\boldsymbol{a \neq b}$ **.**
Cherchons des nombres u et v tels que $\dfrac{1}{(x-a)(x-b)} = \dfrac{v}{x-a} + \dfrac{u}{x-b}$, c'est-à-dire $u(x-a)+v(x-b)=1$. Il vient $u=-v$, $1=-au-bv=(a-b)v$, donc $v=\dfrac{1}{a-b}$ et

$$\frac{P}{(x-a)(x-b)} = \frac{1}{a-b}\left(\frac{P}{x-a} - \frac{P}{x-b}\right)$$

On intègre alors chaque terme de la somme comme en 1.
Plus généralement : si $Q=(x-a)(x-b)(x-c)$, où les nombres a,b,c sont deux à deux différents, on cherche des nombres α,β,γ tels que $\dfrac{1}{Q} = \dfrac{\alpha}{x-a} + \dfrac{\beta}{x-b} + \dfrac{\gamma}{x-c}$.

B) $\boldsymbol{Q(x) = (x-a)(x^2+bx+c)}$ **, où** $\boldsymbol{x^2+bx+c}$ **n'a pas de racine réelle.**
Dans ce cas, on cherche des nombres u, p, q tels que

$$\frac{1}{(x-a)(x^2+bx+c)} = \frac{u}{x-a} + \frac{px+q}{x^2+bx+c}$$

On les calcule en réduisant au même dénominateur et en identifiant. Alors $\dfrac{P}{Q}$ est la somme d'une fraction de la forme $\dfrac{A}{x-a}$ (cas 1) et d'une fraction de la forme $G = \dfrac{B}{x^2+bx+c}$. Pour cette dernière, en divisant comme en 3 le numérateur par le dénominateur, on se ramène à un numérateur de degré inférieur ou égal à 1, comme dans cas 2.

Exemple. Calculons $I = \displaystyle\int \frac{t}{2t^2+t-3}\,dt$. On a $2x^2+x-3 = (x-1)(2x+3)$ et le numérateur est de degré inférieur à 2. On peut donc directement chercher des nombres u et v tels que

$$\frac{x}{(x-1)(2x+3)} = \frac{u}{x-1} + \frac{v}{2x+3}$$

c'est-à-dire $(2x+3)u+(x-1)v=x$. On trouve $2u+v=1$ et $3u-v=0$, d'où $u=1/5$, $v=3/5$ et $I = \dfrac{1}{5}\ln|x-1| + \dfrac{3}{10}\ln|2x+3|$.

6. Intégrales généralisées

Il est parfois naturel de vouloir intégrer sur un intervalle de la forme $[a,+\infty[$: on considère alors les intégrales de a à x et l'on fait tendre x vers l'infini.
Soit f une fonction continue sur $[a,+\infty[$.

Définition

Si les intégrales $\int_a^x f(t)\,dt$ ont une limite finie quand x tend vers $+\infty$, cette limite s'appelle *l'intégrale généralisée de f sur $[a,+\infty[$* et se note $\int_a^{+\infty} f(t)\,dt$.

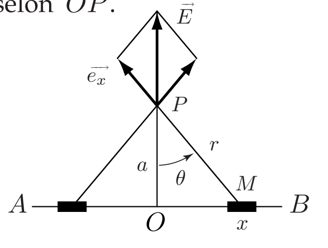
- Dire que f a une intégrale généralisée de valeur v signifie que, quand x tend vers $+\infty$, la fonction $F(x) = \int_a^x f(t)\,dt$ a pour asymptote horizontale la droite d'équation $y = v$.
- Supposons que a' est un nombre supérieur à a ; alors l'intégrale généralisée $\int_{a'}^{+\infty} f(t)\,dt$ existe si et seulement si $\int_a^{+\infty} f(t)\,dt$ existe.

  En effet, les fonctions $x \mapsto \int_a^x f(t)\,dt$ et $x \mapsto \int_{a'}^x f(t)\,dt$ diffèrent de la constante $\int_a^{a'} f(t)\,dt$.
- Si des fonctions f et g ont une intégrale généralisée sur $[a,+\infty[$, alors $f+g$ aussi et $\int_a^{+\infty} f(t)+g(t)\,dt = \int_a^{+\infty} f(t)\,dt + \int_a^{+\infty} g(t)\,dt$.

  En effet, la limite d'une somme de deux fonctions est la somme des limites.

Pour une fonction continue sur $]-\infty,b]$, on définit de même la notion d'intégrale généralisée $\int_{-\infty}^b f(t)\,dt$: c'est la limite quand x tend vers $-\infty$, si elle existe, des intégrales $\int_x^b f(t)\,dt$.

Exemple. Un fil AB chargé électriquement crée en tout point P de l'espace un champ électrique $\vec{E}$ dont la direction est dans le plan PAB. Prenons un point P à égale distance de A et de B ; en appelant O le milieu de AB, P est dans le plan passant par O et orthogonal à AB. Ce plan étant un axe de symétrie de AB, il contient la direction du champ. Finalement, $\vec{E}$ est dirigé selon OP.



Prenons un axe Ox porté par AB. Un petit élément de fil de longueur δx et situé au point M d'abscisse x produit en P le champ $\vec{e_x} = \frac{1}{4\pi\varepsilon_0}\frac{\delta q}{r^2}\vec{u}$, où δq est la charge de l'élément, r la distance MP et $\vec{u}$ le vecteur unitaire dans la direction $\overrightarrow{MP}$ (le nombre ε_0 est la constante diélectrique du milieu ambiant). En supposant que le fil a une densité de charge λ constante, on a $\delta q = \lambda \delta x$. Le champ produit par deux éléments de fil symétriques par rapport au milieu O est dirigé selon $\overrightarrow{OP}$ et son intensité est $\frac{1}{4\pi\varepsilon_0}\frac{\lambda\delta x}{r^2}(2\cos\theta)$, où θ est l'écart angulaire des vecteurs $\overrightarrow{MP}$, $\overrightarrow{OP}$.

Notons a la distance OP. On a $\cos\theta = \frac{a}{r}$, $r^2 = a^2 + x^2$ et $\frac{\cos\theta}{r^2} = \frac{a}{r^3}$, donc le champ électrique en P a pour intensité

$$E = \frac{\lambda a}{2\pi\varepsilon_0}\int_0^{\ell/2}\frac{dx}{(a^2+x^2)^{3/2}}, \quad \text{où } \ell \text{ est la longueur du fil.}$$

La fonction sous le signe intégrale ayant pour primitive $\dfrac{x}{a^2\sqrt{a^2+x^2}}$, on obtient $E = \dfrac{\lambda \ell}{2\pi\varepsilon_0 a\sqrt{4a^2+\ell^2}}$. Faisons tendre la longueur ℓ vers l'infini. Puisque $\sqrt{4a^2+\ell^2} \underset{\ell\to+\infty}{\sim} \ell$, il vient $E \underset{\ell\to+\infty}{\sim} \dfrac{\lambda}{2\pi\varepsilon_0 a}$. Le champ électrique créé à la distance a par un fil de très grande longueur ($\ell \gg a$) a donc pour intensité $\dfrac{\lambda a}{2\pi\varepsilon_0}\displaystyle\int_0^{+\infty} \frac{dx}{(a^2+x^2)^{3/2}} = \frac{\lambda}{2\pi\varepsilon_0 a}$.

Exemples fondamentaux. Soit $a > 0$.

1) Pour la fonction $x \mapsto \dfrac{1}{x}$, l'intégrale $\displaystyle\int_a^x \frac{dt}{t} = \ln x - \ln a$ tend vers $+\infty$ quand x tend vers $+\infty$: l'intégrale généralisée de $1/x$ sur $[a,+\infty[$ n'existe pas.
2) Pour la fonction $x \mapsto \dfrac{1}{x^\alpha}$, où $\alpha \neq 1$, on a

$$\int_a^x \frac{dt}{t^\alpha} = \int_a^x t^{-\alpha}\,dt = \frac{x^{1-\alpha}}{1-\alpha} - \frac{a^{1-\alpha}}{1-\alpha}$$

- Si $\alpha > 1$, alors $1-\alpha < 0$ et $x^{1-\alpha}$ tend vers 0 quand x tend vers $+\infty$. L'intégrale généralisée $\displaystyle\int_a^{+\infty} \frac{dt}{t^\alpha}$ existe et l'on a $\displaystyle\int_a^{+\infty} \frac{dt}{t^\alpha} = -\frac{a^{1-\alpha}}{1-\alpha} = \frac{1}{(\alpha-1)a^{\alpha-1}}$.
- Si $\alpha < 1$, alors $x^{1-\alpha}$ tend vers $+\infty$ quand x tend vers $+\infty$ et l'intégrale généralisée n'existe pas.

Si $a > 0$, l'intégrale généralisée $\displaystyle\int_a^{+\infty} \frac{dt}{t^\alpha}$ existe si et seulement si $\alpha > 1$.

Voici l'outil principal pour étudier l'existence d'une intégrale généralisée lorsqu'on ne peut pas calculer explicitement une primitive.

Théorème de comparaison. *Soient f et g des fonctions continues sur $[a,+\infty[$. Supposons que l'on a $0 \leqslant g(x) \leqslant f(x)$ pour tout x.*

- *Si l'intégrale généralisée de f existe, alors celle de g aussi et $\displaystyle\int_a^{+\infty} g(t)\,dt \leqslant \int_a^{+\infty} f(t)\,dt$.*
- *Si l'intégrale généralisée de g n'existe pas, celle de f non plus.*

Démonstration. Posons $F(x) = \displaystyle\int_a^x f(t)\,dt$ et $G(x) = \displaystyle\int_a^x g(t)\,dt$. Puisque $f(x)$ et $g(x)$ sont positifs ou nuls pour tout x, les fonctions F et G sont croissantes. Supposons que l'intégrale généralisée de f existe. Quand x tend vers l'infini, $F(x)$ a une limite finie, donc est majoré sur $[a,+\infty[$. Puisque $f \leqslant g$, on a $G(x) \leqslant F(x)$, donc $G(x)$ est aussi majoré sur $[a,+\infty[$. La fonction G étant croissante, on en déduit que $G(x)$ a une limite finie quand x tend vers $+\infty$ (voir les propriétés des fonctions monotones, page 264). ∎

Conséquences. *Soient f et g des fonctions à valeurs positives et continues sur $[a,+\infty[$. Supposons que l'on a $g(x) \underset{x\to+\infty}{\ll} f(x)$ ou bien $g(x) \underset{x\to+\infty}{\sim} f(x)$. Si l'intégrale généralisée de f existe, alors celle de g aussi.*

Démonstration. Si $g(x)$ est infiniment petit devant $f(x)$ quand x tend vers $+\infty$, alors pour x assez grand, on a $g(x) \leqslant f(x)$. Si $g(x) \underset{x\to+\infty}{\sim} f(x)$, c'est-à-dire $\lim\limits_{x\to+\infty} \dfrac{g(x)}{f(x)} = 1$, alors pour x assez grand, on a $g(x) \leqslant 2f(x)$ (par exemple). Dans les deux cas, on a donc $g(x) \leqslant 2f(x)$ pour tout $x \geqslant A \geqslant a$. Si l'intégrale $\int_a^{+\infty} f(t)\,dt$ existe, il en va de même de $\int_A^{+\infty} 2f(t)\,dt$. D'après le théorème de comparaison, $\int_A^{+\infty} g(t)\,dt$ existe, donc aussi $\int_a^{+\infty} g(t)\,dt$. ■

Pour voir si l'intégrale généralisée d'une fonction positive f existe, on peut remplacer $f(x)$ par un équivalent en $+\infty$.

Exemples

1) On a $\dfrac{\sqrt{x}}{x^2+1} \underset{x\to+\infty}{\sim} \dfrac{1}{x^{3/2}}$, donc l'intégrale généralisée $\int_0^{+\infty} \dfrac{\sqrt{t}}{t^2+1}\,dt$ existe.

2) Puisque $x^2e^{-x^2}$ tend vers 0 quand x tend vers $+\infty$, on a $e^{-x^2} \underset{x\to+\infty}{\ll} (1/x^2)$ et comme l'intégrale généralisée $\int_1^{+\infty} (1/t^2)\,dt$ existe, il en va de même de $\int_0^{+\infty} e^{-t^2}\,dt$. La fonction e^{-x^2} étant paire, on a aussi $\int_{-\infty}^0 e^{-t^2}\,dt = \int_0^{+\infty} e^{-t^2}\,dt$.

3) La fonction $1/(x\ln x)$ a pour primitive $\ln(\ln x)$ qui tend vers $+\infty$ quand x tend vers l'infini : $1/(x\ln x)$ n'a donc pas d'intégrale généralisée sur $[2,+\infty[$, bien que cette fonction soit infiniment petite devant $1/x$.

4) L'intégrale généralisée $\int_1^{+\infty} \dfrac{(\sin t)^2}{t^\alpha}\,dt$ existe si $\alpha > 1$, car $\dfrac{(\sin t)^2}{t^\alpha} \leqslant \dfrac{1}{t^\alpha}$.

Cas d'une fonction qui ne garde pas un signe constant. Nous allons voir que l'intégrale généralisée existe pourvu que la fonction ne soit pas trop grande en valeur absolue.

Proposition. *Si l'on a $\left|f(x)\right| \leqslant g(x)$ pour tout $x \geqslant a$ et si l'intégrale généralisée de g existe, alors celle de f aussi.*

Démonstration. Posons $h(x) = \left|f(x)\right| - f(x)$ de sorte que l'on a $h(x) \geqslant 0$. Puisque $-f(x) \leqslant \left|f(x)\right|$, on a $0 \leqslant h(x) \leqslant \left|2f(x)\right| \leqslant 2g(x)$. L'intégrale généralisée de $2g(x)$ existe, donc aussi celle de $h(x)$, d'après le théorème de comparaison. La fonction $f(x) = \left|f(x)\right| - h(x)$ est différence de deux fonctions ayant une intégrale généralisée, donc f a une intégrale généralisée. ■

Exemple. Pour tout x, on a $\left|\dfrac{\sin x}{x^2+1}\right| \leqslant \dfrac{1}{x^2+1}$. Puisque $\int_0^{+\infty} \dfrac{dt}{t^2+1} = \lim\limits_{x\to+\infty} \arctan x = \dfrac{\pi}{2}$, on en déduit que l'intégrale généralisée $\int_0^{+\infty} \dfrac{\sin t}{t^2+1}\,dt$ existe et que sa valeur est moindre que $\dfrac{\pi}{2}$. On a aussi $\int_{-\infty}^0 \dfrac{\sin t}{t^2+1}\,dt = -\int_0^{+\infty} \dfrac{\sin t}{t^2+1}\,dt$, car $x \mapsto \dfrac{\sin x}{x^2+1}$ est impaire.

Fonction tendant vers l'infini en un point

Exemples fondamentaux. Soit α un nombre positif. Quand x tend vers 0, $1/x^\alpha$ tend vers l'infini, mais étudions la limite des intégrales $F(x) = \int_x^1 \frac{dt}{t^\alpha}$. On a

$$F(x) = \begin{cases} \dfrac{1 - x^{1-\alpha}}{1-\alpha} & \text{si } \alpha \neq 1 \\ -\ln x & \text{si } \alpha = 1 \end{cases}$$

- Si $\alpha < 1$, alors $x^{1-\alpha}$ tend vers 0 quand x tend vers 0 et $F(x)$ a pour limite $1/(1-\alpha)$. On peut alors définir l'intégrale généralisée de la fonction $1/x^\alpha$ sur $]0,1]$ en posant $\int_0^1 \frac{dt}{t^\alpha} = \lim\limits_{x \to 0} F(x) = \frac{1}{1-\alpha}$.
- Si $\alpha \geqslant 1$, alors $\lim\limits_{x \to 0} F(x) = +\infty$ et l'intégrale généralisée sur $]0,1]$ n'existe pas.

Définition

Soit f une fonction continue sur l'intervalle semi-ouvert $[a,b[$. Si les intégrales $\int_a^x f(t)\,dt$ ont une limite finie quand x tend vers b, on note cette limite $\int_a^b f(t)\,dt$ et on l'appelle *l'intégrale généralisée de f sur $[a,b[$.*

Le théorème de comparaison, ses conséquences et la proposition précédente sont encore vraies dans ce cadre, en remplaçant « x tend vers l'infini » par « x tend vers b ». De manière analogue, on définit la notion d'intégrale généralisée pour une fonction continue sur un intervalle semi-ouvert $]a,b]$, comme dans l'exemple précédent.

Exemple 1. La période T d'un pendule simple de longueur ℓ dépend de l'angle θ_0 dont on l'a écarté par rapport à la verticale ($0 < \theta_0 < \pi/2$) :

$$T = 4\sqrt{\frac{\ell}{2g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos\theta - \cos\theta_0}}$$

Dans la fonction sous le signe intégrale, le dénominateur s'annule si $\theta = \theta_0$, donc la limite est $+\infty$ quand θ tend vers θ_0 : l'intégrale est une intégrale généralisée. Pour vérifier son existence, cherchons un équivalent de $\sqrt{\cos\theta - \cos\theta_0}$ quand θ tend vers θ_0. Puisque $\cos'\theta_0 = -\sin\theta_0$, on sait que $\cos\theta - \cos\theta_0 \underset{\theta \to \theta_0}{\sim} (\theta - \theta_0)(-\sin\theta_0)$, donc

$$\sqrt{\cos\theta - \cos\theta_0} \underset{\theta \to \theta_0}{\sim} (\sqrt{\sin\theta_0})\sqrt{\theta_0 - \theta}$$

et la fonction sous le signe intégrale est équivalente à $\frac{1/\sqrt{\sin\theta_0}}{(\theta_0 - \theta)^{1/2}}$. La fonction $\frac{1}{(\theta_0 - \theta)^{1/2}}$ a une intégrale généralisée sur $[0, \theta_0[$, car l'exposant $1/2$ est strictement inférieur à 1. On en déduit que l'intégrale exprimant la période T est bien définie.

Exemple 2. L'intégrale généralisée $\int_{-1}^{1} \frac{dt}{\sqrt{1-t^2}}$ vaut π : en effet, d'après le tableau des primitives page 298, $\int \frac{dt}{\sqrt{1-t^2}} = \operatorname{Arc}\sin x$, donc $\int_{-1}^{1} \frac{dt}{\sqrt{1-t^2}} = 2\operatorname{Arc}\sin 1 = \pi$.

Exemple 3. Posons $f(x)=1/\sin x$. Quand x tend vers 0, $f(x)$ est équivalent à $1/x$ et l'intégrale généralisée de $1/x$ n'existe pas sur $]0,\pi/2]$, donc celle de $f(x)$ non plus.

7. Application aux probabilités

7.1 Fonction de répartition et densité

Soit X une variable aléatoire prenant des valeurs réelles.

Définition

Pour tout nombre réel x, posons $F(x)=P(X \leqslant x)$, probabilité pour que la valeur de X soit inférieure ou égale à x. La fonction F s'appelle la *fonction de répartition* de X.

La probabilité pour que la valeur de X soit dans l'intervalle $]a,b]$ est

$$P(a < X \leqslant b) = F(b) - F(a)$$

Exemple. Supposons que X est la somme des points donnés par un lancer de deux dés. Le nombre de façons d'obtenir k points dépend de la valeur de k : si k est entre 2 et 7, il y a $k-1$ façons, sinon il y en a $13-k$. On en déduit la probabilité p_k d'obtenir k points :

k	2	3	4	5	6	7	8	9	10	11	12
p_k	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

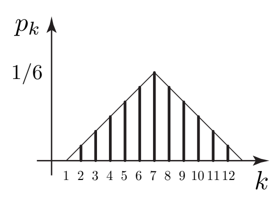
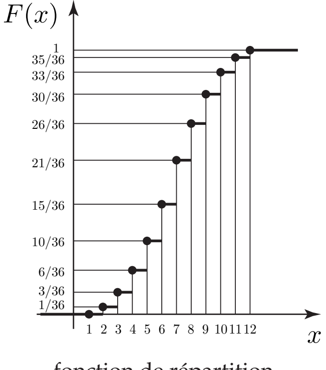


diagramme en batons



fonction de répartition



Cette fonction de répartition est en escalier.

Propriétés d'une fonction de répartition

a) La fonction de répartition est croissante : en effet, si $x \leqslant y$ et si l'événement $X \leqslant x$ est réalisé, alors l'événement $X \leqslant y$ aussi.

b) Pour tout x, on a $0 \leqslant F(x) \leqslant 1$, car $F(x)$ est une probabilité.

c) $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to +\infty} F(x) = 1$.

d) Si des nombres $x_n \geqslant x$ tendent vers x en décroissant, alors $F(x_n)$ tend vers $F(x)$.

Définition

S'il existe une fonction f à valeurs positives telle que $F(x) = \int_{-\infty}^{x} f(t)\,dt$ pour tout x, on dit que f est une *densité* de la variable aléatoire X.

- On a $\int_{-\infty}^{+\infty} f(t)\,dt = \lim_{x \to +\infty} F(x) = 1$.
- La densité d'une variable aléatoire détermine entièrement sa loi de probabilité.
- Si la densité est continue, alors F est dérivable et $F'(x) = f(x)$.

Choisissons en effet un nombre a quelconque. On a

$$F(x) - F(a) = \int_{-\infty}^{x} f(t)\,dt - \int_{-\infty}^{a} f(t)\,dt = \int_{a}^{x} f(t)\,dt\,;$$

cela montre que si f est continue, la fonction $x \mapsto F(x) - F(a)$ a pour dérivée $f(x)$, donc aussi la fonction F.

Définitions

Soit X une variable aléatoire de densité f.

- *L'espérance* de X est $E(X) = \int_{-\infty}^{+\infty} t f(t)\,dt$, si cette intégrale généralisée existe.
- La *variance* de X est l'espérance de la variable $X^2 - E(X)^2$, c'est-à-dire le nombre $V(X) = \int_{-\infty}^{+\infty} \left[t^2 - E(X)^2\right] f(t)\,dt$, si cette intégrale généralisée existe. *L'écart-type* est $\sqrt{V(X)}$.

Propriétés

i) *Si a et b sont des nombres, alors $E(aX + b) = aE(X) + b$.*

ii) *Si des variables aléatoires X et Y ont une espérance, alors $E(X + Y) = E(X) + E(Y)$.*

iii) $V(X) = E\left[\left(X - E(X)\right)^2\right]$.

Justification. Pour $a \neq 0$, l'inégalité $aX+b \leqslant x$ s'écrit $X \leqslant (x-b)/a$; si F est la loi de répartition de X, alors la loi de $aX+b$ est $F\big((x-b)/a\big)$ et sa densité est $g(x) = (1/a) f\big((x-b)/a\big)$. On a donc

$$\begin{aligned} E(aX + b) &= \frac{1}{a} \int_{-\infty}^{+\infty} t f\left(\frac{t-b}{a}\right) dt \\ &= \frac{1}{a} \int_{-\infty}^{+\infty} (au + b) f(u)\, a du \quad \text{(changement de variable } t = au + b) \\ &= a \int_{-\infty}^{+\infty} u f(u)\,du + b \int_{-\infty}^{+\infty} f(u)\,du = aE(X) + b\,. \end{aligned}$$

Montrons maintenant (iii). En posant $E(X)=m$, on a $(X-m)^2 = X^2 - 2mX + m^2$, donc en prenant l'espérance, il vient $E\big[(X-m)^2\big] = E(X^2) - 2mE(X) + m^2 = E(X^2) - m^2 = E(X^2 - m^2) = V(X)$. ■

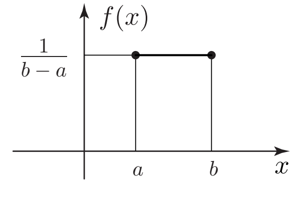
7.2 Des exemples de lois de probabilité

La loi uniforme

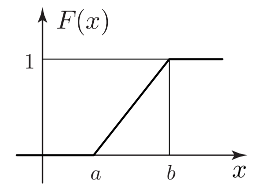
Soient a et b des nombres tels que $a < b$ et soit f la fonction définie par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leqslant x \leqslant b \\ 0 & \text{sinon} \end{cases}$$

Une variable aléatoire X de densité f est dite *uniforme*. Son espérance est $\int_{-\infty}^{+\infty} f(t)\,dt = \int_a^b \frac{t}{b-a}\,dt = \frac{1}{b-a}\,\frac{b^2-a^2}{2}$, donc $E(X) = \frac{a+b}{2}$.



densité uniforme



fonction de répartition

La loi exponentielle

Exemple. L'instant de désintégration d'un atome d'élément radioactif est un événement aléatoire. Si un atome ne s'est pas désintégré à un instant donné t, la probabilité qu'il se désintègre au cours d'un intervalle de temps à partir de cet instant ne dépend que de la longueur de l'intervalle. Appelons $F(t)$ la probabilité de désintégration avant l'instant t.

Si X est la variable aléatoire « instant de désintégration », on a donc $P(X<t)=F(t)$. La probabilité de désintégration au cours d'un petit intervalle de temps $[t, t+\delta t]$ est proportionnelle à δt et à la probabilité $1-F(t)$ de non-désintégration avant l'instant t. On a ainsi

$$F(t+\delta t) - F(t) = \big[1 - F(t)\big]\lambda\delta t\,, \quad \text{avec } \lambda \text{ constant.}$$

En faisant tendre δt vers 0 et en posant $y = F(t)$, il vient

$$dy = (1-y)\lambda dt \quad , \quad \frac{dy}{1-y} = \lambda dt \ , \quad \text{et en intégrant :}$$

$$\int \frac{dy}{1-y} = -\ln(1-y) = \lambda t - c, \quad \text{où } c \text{ est une constante.}$$

Prenons comme instant initial $t = 0$, de sorte que pour $t = 0$, on a $y = F(0) = 0$. Alors $c = 0$, $1 - y = e^{-\lambda t}$ et

$$F(t) = 1 - e^{-\lambda t}, \quad \text{pour tout } t > 0.$$

C'est la loi de répartition de X pour $t > 0$. Si $t \leqslant 0$, on a bien entendu $F(t) = 0$.

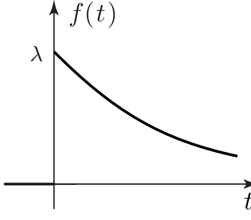
La densité de X est $f(t) = \begin{cases} 0 & \text{si } t \leqslant 0 \\ F'(t) = \lambda e^{-\lambda t} & \text{si } t > 0 \end{cases}$

L'espérance de X est

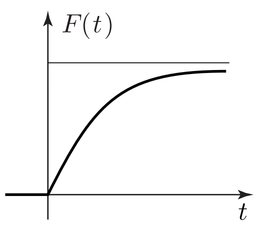
$$E(X) = \int_0^{+\infty} \lambda t e^{-\lambda t}\, dt = \left[-\left(t + \frac{1}{\lambda}\right) e^{-\lambda t} \right]_0^{+\infty} = \frac{1}{\lambda}$$

On a $\displaystyle\int_0^{+\infty} \lambda t^2 e^{-\lambda t}\, dt = \left[-\left(t^2 + \frac{2}{\lambda} t + \frac{2}{\lambda^2}\right) e^{-\lambda t} \right]_0^{+\infty} = \frac{2}{\lambda^2}$, donc la variance est

$$V(X) = \int_0^{+\infty} \left(t^2 - \frac{1}{\lambda^2}\right) \lambda e^{-\lambda t}\, dt = \frac{2}{\lambda^2} - \frac{1}{\lambda} \int_0^{+\infty} e^{-\lambda t}\, dt = \frac{1}{\lambda^2}$$



densité exponentielle



fonction de répartition

> **Définition**
> On dit qu'une variable aléatoire suit une *loi de probabilité exponentielle* si sa densité est de la forme $f(t) = \begin{cases} 0 & \text{si } t < 0 \\ \lambda e^{-\lambda t} & \text{si } t \geqslant 0 \end{cases}$, où λ est un nombre positif.

En raison de son importance pratique, nous allons étudier plus en détail la loi de Laplace-Gauss, dite aussi loi normale.

7.3 La loi normale

Une approche de la loi normale

Effectuons n épreuves d'un événement aléatoire qui suit une loi binomiale de probabilité p, avec $0 < p < 1$. La probabilité que k événements se réalisent est

$$W_k = \binom{n}{k} p^k q^{n-k}\ , \quad \text{où l'on a posé } q = 1 - p.$$

L'espérance du nombre de réalisations est $E = np$ et la variance est $V = npq$ (pages 67-68).

Nous allons chercher l'ordre de grandeur de W_k quand k et n deviennent grands d'une manière telle que k reste compris entre $E+\alpha\sqrt{V}$ et $E+\beta\sqrt{V}$, où α et β sont des nombres positifs fixés tels que $\alpha<\beta$.

Ces conditions signifient que la différence $k-E$ reste comprise entre $\alpha\sqrt{V}$ et $\beta\sqrt{V}$.

Nous utiliserons pour cela la formule de Stirling : $n! \underset{n\to+\infty}{\sim} n^n e^{-n}\sqrt{2\pi n}$.

Posons $z=k-E=k-np$, donc $k=np+z$ et $n-k=nq-z$. On a

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}\sim\frac{n^n}{(np{+}z)^{np+z}(nq{-}z)^{nq-z}}\,\frac{e^{-n}}{e^{-(np+z)}e^{-(nq-z)}}\,\frac{\sqrt{2\pi n}}{\sqrt{2\pi(np{+}z)}\sqrt{2\pi(nq{-}z)}}$$

Examinons ce produit : puisque $(np+z)+(nq-z)=n$, la fraction du milieu vaut 1 et pour la même raison, on a $n^n=n^{np+z}n^{nq-z}$, donc

$$n^np^kq^{n-k}=n^{np+z}n^{nq-z}p^{np+z}q^{nq-z}=(np)^{np+z}(nq)^{nq-z}$$

Par suite

$$W_k\sim\sqrt{\frac{n}{2\pi(np{+}z)(nq{-}z)}}\;\frac{(np)^{np+z}}{(np{+}z)^{np+z}}\;\frac{(nq)^{nq-z}}{(nq{-}z)^{nq-z}}$$

$$(*)\qquad W_k\sim\sqrt{\frac{n}{2\pi(np{+}z)(nq{-}z)}}\;\left(1-\frac{z}{np{+}z}\right)^{np+z}\;\left(1+\frac{z}{nq{-}z}\right)^{nq-z}$$

Posons $x=\dfrac{z}{\sqrt{npq}}$, quantité qui par hypothèse reste bornée quand n tend vers l'infini. Puisque z/np et z/nq tendent vers 0, $(np+z)(nq-z)=n^2pq\left(1+\frac{z}{np}\right)\left(1-\frac{z}{nq}\right)$ est équivalent à n^2pq, donc le premier facteur dans $(*)$ est équivalent à $\dfrac{1}{\sqrt{2\pi npq}}$.

Remarquons que $\dfrac{np}{z}=\dfrac{np}{x\sqrt{npq}}$ est de l'ordre de $\sqrt{n}$, donc $\dfrac{z}{np+z}=\dfrac{1}{(np/z)+1}$ est de l'ordre de $1/\sqrt{n}$ et tend vers 0 ; de même pour $\dfrac{z}{nq{-}z}$.

Écrivons un développement limité de $(np+z)\ln\left(1-\dfrac{z}{np+z}\right)$. Puisque $\ln(1+u)=u-\dfrac{u^2}{2}+o(u^2)$ quand u tend vers 0, il vient

$$(np+z)\ln\left(1-\frac{z}{np+z}\right)=-z-\frac{z^2}{2(np+z)}+a_n\,,$$

où a_n est infiniment petit devant $\dfrac{z^2}{np+z}=x\sqrt{npq}\dfrac{z}{np+z}$, donc tend vers 0.

De même, $(nq-z)\ln\left(1+\dfrac{z}{nq-z}\right)=z-\dfrac{z^2}{2(nq-z)}+b_n$, avec b_n tendant vers 0.

En ajoutant et en posant $c_n=a_n+b_n$, on obtient

$$(np+z)\ln\left(1-\frac{z}{np+z}\right)+(nq-z)\ln\left(1-\frac{z}{nq-z}\right)=-\frac{z^2}{2}\,\frac{n}{(np{+}z)(nq{-}z)}+c_n$$

$$=-\frac{x^2}{2}npq\frac{n}{(np{+}z)(nq{-}z)}+c_n=-\frac{x^2}{2}\,\frac{1}{\big[1+(z/np)\big]\big[1-(z/nq)\big]}+c_n$$

quantité qui est de l'ordre de $-x^2/2$. On en déduit que W_k est équivalent à $\dfrac{1}{\sqrt{2\pi npq}}e^{-x^2/2}$, c'est-à-dire

$$W_k\sim\frac{1}{\sqrt{2\pi}\sqrt{npq}}\;\exp\left[-\frac{(k-np)^2}{2npq}\right]$$

Dans nos conditions de passage à la limite, la probabilité binomiale $\binom{n}{k}p^k q^{n-k}$ est donc approchée par la valeur en $x = k$ de la fonction

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sqrt{V}} \exp\left[-\frac{(x-E)^2}{2V}\right], \quad \text{où } E = np \text{ et } V = npq.$$

La probabilité pour que le nombre k de réalisations de l'événement soit compris entre $E + \alpha\sqrt{V}$ et $E + \beta\sqrt{V}$ est $W = \sum_{\alpha \leqslant x_k \leqslant \beta} W_k$, où l'on a posé $x_k = \frac{k-np}{\sqrt{npq}}$, c'est-à-dire approximativement

$$W \sim \frac{1}{\sqrt{2\pi}\sqrt{npq}} \sum_{\alpha \leqslant x_k \leqslant \beta} e^{-(x_k)^2/2} = \frac{1}{\sqrt{2\pi}} \sum_{\alpha \leqslant x_k \leqslant \beta} (x_{k+1} - x_k) e^{-(x_k)^2/2}$$

car $x_{k+1} - x_k = \frac{1}{\sqrt{npq}}$. Puisque $x_{k+1} - x_k$ tend vers 0, la somme tend, par définition de l'intégrale, vers $\int_\alpha^\beta e^{-t^2/2}\,dt$ (page 286).

La probabilité pour qu'au cours de n épreuves, le nombre de réalisations soit compris entre $np + \alpha\sqrt{npq}$ et $np + \beta\sqrt{npq}$ tend vers $\frac{1}{\sqrt{2\pi}} \int_\alpha^\beta e^{-t^2/2}\,dt$.

Application. Dans la pratique, on peut utiliser cette approximation pour évaluer une probabilité binomiale lorsque n est grand et que p n'est ni trop grand, ni trop petit. Dans ces conditions, la probabilité pour qu'au cours de n épreuves, le nombre de réalisations soit compris entre k_1 et k_2, est à peu près

$$\frac{1}{\sqrt{2\pi}} \int_a^b \exp(-t^2/2)\,dt, \quad \text{où } a = \frac{k_1 - np}{\sqrt{npq}} \text{ et } b = \frac{k_2 - np}{\sqrt{npq}}$$

La fonction de Gauss

La fonction $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ est une densité de probabilité.

En effet, c'est une fonction positive, ses intégrales généralisées $\int_{-\infty}^{x} \varphi(t)\,dt$ et $\int_{-\infty}^{+\infty} \varphi(t)\,dt$ existent (d'après l'exemple 2 page 326) et enfin, nous montrerons au prochain chapitre que l'on a

$$\int_{-\infty}^{+\infty} \varphi(t)\,dt = 1$$

Définition

La fonction de répartition $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt$ s'appelle la *fonction de Gauss*.

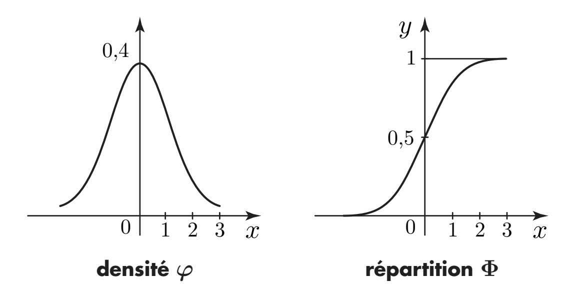
La fonction φ étant paire, on a $\int_{-\infty}^{-x} \varphi(t)\,dt = \int_{x}^{+\infty} \varphi(t)\,dt = 1 - \int_{-\infty}^{x} \varphi(t)\,dt$, c'est-à-dire

$$\Phi(x) + \Phi(-x) = 1$$

Puisque $te^{-t^2/2}$ est une fonction impaire, l'intégrale $\int_{-\infty}^{+\infty} te^{-t^2/2}\,dt$ vaut 0 :

si une variable aléatoire a pour fonction de répartition Φ, son espérance est nulle.

Les valeurs $\Phi(x)$ ne s'expriment pas au moyen des fonctions usuelles, mais on peut en faire un calcul approché au moyen d'une table numérique (voir page 577).



densité φ **répartition Φ**

Exemple. On jette un dé 1200 fois. Quelle est la probabilité pour obtenir entre 180 et 210 fois le nombre 1 ?
La probabilité est

$$P = \sum_{k=180}^{210} \binom{1200}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{1200-k}.$$

On peut l'approcher par $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_a^b \exp(-t^2/2)\,dt$, où $a = \dfrac{180-200}{\sqrt{1200 \times (5/36)}} \simeq -1{,}55$ et $b = \dfrac{210-200}{\sqrt{1200 \times (5/36)}} \simeq 0{,}774$. On trouve ainsi $P \simeq \Phi(b) - \Phi(a) \simeq 0{,}780 - 0{,}06 = 0{,}72$.
La valeur exacte est proche de $0{,}738$.

La loi normale

Généralisons la densité de Gauss pour rendre compte, en particulier, de variables aléatoires dont la moyenne n'est pas nulle.

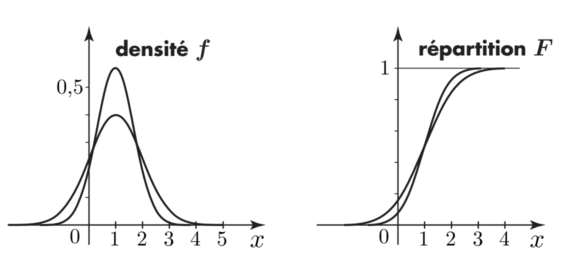
Définition

Soient m et σ des nombres positifs. On dit qu'une variable aléatoire suit une loi *normale* de paramètres m et σ (en abrégé $\mathcal{N}(m,\sigma)$), si sa densité est la fonction

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$$

La fonction de répartition de X est alors $F(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} \exp\left[-\frac{(t-m)^2}{2\sigma^2}\right] dt$.

On a $F(x) = \Phi\left(\dfrac{x-m}{\sigma}\right)$ (changement de variable $\sigma u = t - m$ dans l'intégrale). Les figures ci-contre montrent la densité f et la fonction de répartition F de la loi normale lorsque $m=1$, pour $\sigma=1$ et $\sigma=0{,}7$ (comparer aux courbes page 67).



Dans une série de mesures, les erreurs sont souvent distribuées selon la loi normale : c'est pourquoi on l'appelle quelquefois « loi des erreurs ».

Espérance et variance. Si une variable aléatoire X suit la loi normale $\mathcal{N}(m, \sigma)$, alors

$$E(X) = m \quad \text{et} \quad V(X) = \sigma^2$$

Ces paramètres permettent de mesurer la probabilité pour que la valeur de X reste dans un intervalle donné autour de la moyenne. Pour tout nombre $\alpha > 0$, on a en effet

$$P\big(|X-m| < \alpha\sigma\big) = F(m+\alpha\sigma) - F(m-\alpha\sigma) = \Phi(\alpha) - \Phi(-\alpha) = 2\Phi(\alpha) - 1$$

Par exemple, $P\big(|X-m| < \sigma\big) = 2\Phi(1) - 1 \simeq 0{,}68$ et $P\big(|X-m| < 2\sigma\big) = 2\Phi(2) - 1 \simeq 0{,}95$.

Nous avons montré comment, pour n grand, la loi binomiale tend, au sens des probabilités, vers la loi normale de Gauss. Ce résultat est encore vrai dans un contexte beaucoup plus général et constitue une propriété remarquable de la loi normale. Pour l'énoncer, introduisons la notion de variables aléatoires indépendantes.

Définition

Soient X et Y des variables aléatoires à valeurs réelles. On dit que X et Y sont *indépendantes* si pour tous intervalles $]a, b]$ et $]c, d]$, les événements $a < X \leqslant b$ et $c < Y \leqslant d$ sont indépendants, c'est-à-dire si

$$P\big[(a < X \leqslant b) \text{ et } (c < Y \leqslant d)\big] = P(a < X \leqslant b)\, P(c < Y \leqslant d)$$

Si l'ensemble des événements possibles est fini, les variables prennent un nombre fini de valeurs $x_1, x_2, \ldots, x_p$ pour X, $y_1, y_2, \ldots, y_n$ pour Y. Dans ce cas, X et Y sont indépendantes si et seulement si pour tout couple (i, j), la probabilité de l'événement $(X = x_i$ et $Y = y_j)$ est le produit $P(X = x_i)P(Y = y_j)$ des probabilités.

Théorème de la limite centrée. *Soit (X_n) une suite de variables aléatoires deux à deux indépendantes, de même loi, d'espérance m et de variance σ^2. Posons $\overline{X}_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. Alors pour tous nombres $a < b$, $P\left(\dfrac{\sigma a}{\sqrt{n}} < \overline{X}_n - m \leqslant \dfrac{\sigma b}{\sqrt{n}}\right)$ tend vers $\Phi(b) - \Phi(a)$ quand n tend vers l'infini.*

Supposons que chaque X_i représente le résultat d'une seule épreuve de probabilité p telle que $0 < p < 1$: on a $X_i = 1$ en cas de succès, $X_i = 0$ en cas d'échec. L'espérance de X_i est $m = p$ et sa variance est $\sigma^2 = pq$, où $q = 1 - p$. La valeur de $X_1 + X_2 + \cdots + X_n$ est le nombre k de succès en n épreuves. L'encadrement $\sigma a/\sqrt{n} < k/n - m \leqslant \sigma b/\sqrt{n}$ s'écrit $np + a\sqrt{npq} < k \leqslant np + b\sqrt{npq}$ (en multipliant par n) : l'étude menée en introduction montre donc le théorème dans ce cas particulier.

Comme dans l'exemple précédent, ce théorème permet de calculer de manière approchée des probabilités de la moyenne $\overline{X}_n$ pour n grand.

Application au calcul d'un intervalle de confiance

Au terme d'un processus de fabrication, on effectue un contrôle de qualité des pièces produites : sur un lot de 200 pièces, 24 sont défectueuses.

Premier problème. On veut estimer la proportion p de pièces défectueuses avec un un indice de certitude de 0,95.

Définissons la variable aléatoire X qui prend la valeur 1 avec la probabilité p (cas d'une mauvaise pièce) et la valeur 0 avec la probabilité $1-p$ (cas d'une bonne pièce).

- L'espérance de X est $m = 1 \times P(X=1) + 0 \times P(X=0) = p$
- et sa variance est

$$\sigma^2 = (0-m)^2 P(X=0) + (1-m)^2 P(X=1) = p^2(1-p)+(1-p)^2 p = p(1-p)$$

Notre échantillon est représenté par une suite $X_1, \ldots, X_n$ $(n=200)$ de variables aléatoires de même loi que X. Faisons l'hypothèse que ces variables sont indépendantes (bonne qualité de l'échantillon) et posons $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$: c'est la moyenne empirique des observations, donc $\overline{X}_n = 24/200 = 0{,}12$.

D'après le théorème de la limite centrée, la probabilité

(1) $$P\Big[\, \big|\overline{X}_n - p\big| < \frac{2\sigma}{\sqrt{n}} \,\Big]$$

vaut à peu près 0,95. En estimant l'écart-type σ par $\sqrt{\overline{X}_n(1-\overline{X}_n)} \simeq 0{,}325$, on obtient l'approximation $P\big[\,|p-0{,}12| < 0{,}046\,\big] \simeq 0{,}95$, ou encore

$$P\big[\,0{,}07 < p < 0{,}16\,\big] \simeq 0{,}95$$

Une pièce donnée a donc une probabilité comprise entre 0,07 et 0,16 d'être défectueuse, avec un indice de certitude de 95%. L'intervalle $[0{,}07\,,\,0{,}16]$ s'appelle un *intervalle de confiance pour l'estimation de p au risque* 0,05.

Si l'on veut un intervalle de confiance au risque 0,01, il faut remplacer dans (1) le coefficient 2 par le nombre a tel que $2\Phi(a) - 1 = 0{,}99$, soit environ $a = 2{,}58$ (voir la table page 577) ; on obtient un intervalle plus grand.

Second problème. Combien suffit-il de contrôler de pièces pour avoir, avec un risque 0,05, un intervalle de confiance centré de longueur 0,04 ?

On cherche n pour que, dans (1), l'amplitude $\dfrac{2\sqrt{p(1-p)}}{\sqrt{n}}$ de la fourchette soit inférieure à 0,04. Comme on ne connaît pas bien p, majorons le produit $p(1-p)$ par $1/4$ (c'est le maximum de la fonction $t(1-t)$ pour $0 \leqslant t \leqslant 1$). On trouve ainsi $\dfrac{1}{\sqrt{n}} \leqslant 0{,}04$, d'où $n \geqslant (25)^2 = 625$.

Exercices

@ **1.** On considère la parabole d'équation $y = x^2$ et le point $A = (0, a)$, où $a > 0$. Écrire la distance AM quand $M = (x, y)$ est un point de la parabole. Étudier les variations de cette distance en fonction de x. Quel est le minimum de la distance AM quand le point M parcourt la parabole ? (discuter selon la valeur de a).

@ **2. Un exemple d'attraction par un point fixe.** Dans l'exercice **4** du chapitre précédent, nous avons vu que si a est un nombre strictement compris entre entre 0 et 1, la fonction $f(x) = e^{-ax}$ a un point fixe $s(a) > 0$.

a) Montrer que l'on a $\left|f'(x)\right| < a$ pour tout $x > 0$. En déduire que $s(a)$ est un point fixe attractif de f.

b) Pour tout entier $n \geqslant 0$, définissons les nombres u_n en posant $u_0 = 1$ et $u_{n+1} = e^{-au_n}$. Montrer que l'on a $\left|u_n - s(a)\right| \leqslant a^n\big(1 - s(a)\big)$. Calculer à 10^{-3} près la solution de l'équation $2x = e^{-x}$ (faire un changement d'inconnues).

@ **3. Une étude de point fixe.** Soit a un nombre strictement positif. Pour tout $x \geqslant 0$, posons $f(x) = e^{a(x-1)}$ et définissons les nombres u_n par $u_0 = 0$ et $u_{n+1} = f(u_n)$.

a) Dessiner le graphe de f. Montrer que l'on a $0 \leqslant u_n \leqslant 1$ et $u_{n+1} \geqslant u_n$ pour tout n (raisonner par récurrence). En déduire que les u_n ont une limite inférieure ou égale à 1. Montrer que si $a = 1$, cette limite vaut 1.

b) On suppose $0 < a < 1$. Montrer que si $0 \leqslant x < 1$, alors $f'(x) \leqslant a$; en déduire $|u_n - 1| \leqslant a^n |u_0 - 1|$ et $\lim u_n = 1$.

c) On suppose désormais $a > 1$.

(i) Montrer que la fonction f' est strictement croissante et tend vers $+\infty$ quand x tend vers $+\infty$. En déduire qu'il y a un unique nombre $\alpha \geqslant 0$ tel que $f'(\alpha) = 1$. Calculer $f'(0)$ et $f'(1)$ et montrer que l'on a $0 < \alpha < 1$.

(ii) Montrer que la fonction $g(x) = f(x) - x$ est décroissante sur $[0, \alpha]$, croissante sur $[\alpha, +\infty[$, qu'elle vaut 0 en $x = 1$ et que son minimum en α est négatif. En déduire que l'équation $f(x) = x$ a deux solutions : l'une, qu'on note $r(a)$, entre 0 et α, l'autre supérieure à α. On a donc $0 < r(a) < 1$. Montrer que les nombres u_n sont tous entre 0 et $r(a)$ et que $\lim u_n = r(a)$.

(iii) Montrer que la fonction $\varphi(x) = xe^{-x}$ atteint son maximum $1/e$ en $x = 1$. Montrer que le nombre $b = ar(a)$ vérifie $\varphi(b) = \varphi(a)$ et que $b < a$. En déduire que $b < 1$ et que $r(a)$ est entre 0 et $1/a$.

(iv) En admettant que la fonction $t \mapsto r(t)$ est dérivable et en utilisant la relation $f\big(r(a)\big) = r(a)$, montrer que $r(t)$ est solution de l'équation différentielle $r' = \dfrac{r(1-r)}{tr-1}$. En déduire que la fonction $r(t)$ est décroissante et qu'elle tend vers 0 quand t tend vers $+\infty$.

(v) Posons $m = e\varphi(a)$. Montrer que $m < 1$, que $f'(1/a) = m$ et que pour tout $x \in [0, 1/a]$, on a $0 \leqslant f'(x) \leqslant m$. En déduire que pour tout n, on a $|u_n - r(a)| \leqslant m^n \, |u_0 - r(a)| = r(a)m^n$.

4. Estimation des paramètres d'une diffusion. Comme dans l'exemple page 305, supposons qu'après injection, la quantité de substance diffusée varie au cours du temps selon la formule $q(t) = ae^{-\alpha t} + be^{-\beta t}$, où le temps t est compté en heures à partir d'un certain instant origine.

a) Voici des mesures de $\ln q(t)$:

t	2	3	4	5	6
$\ln q(t)$	$-0{,}57$	$-0{,}95$	$-1{,}37$	$-1{,}75$	$-2{,}16$

Montrer qu'on peut prendre comme estimations $a = 1{,}27$ et $\alpha = 0{,}4$ (chercher une droite de régression, comme page 211).

b) Posons $r(t) = 1{,}27\, e^{-0{,}4\, t} - q(t)$. Voici les mesures du logarithme de $r(t)$:

t	2	3	4	5	6
$\ln r(t)$	$-5{,}1$	$-7{,}4$	$-9{,}75$	-12	$-14{,}5$

Montrer que l'on peut adopter la formule $q(t) = 1{,}27\, e^{-0{,}4\, t} - 0{,}85\, e^{-2{,}4\, t}$.

c) Comment varie $q(t)$ pour $0 \leqslant t \leqslant 6$? Quel est son maximum ? Dessiner le graphe de cette fonction.

5. Sur un même dessin, représenter l'allure du graphe des fonctions suivantes au voisinage de $x = 1$; tenir compte de la position du graphe par rapport à la tangente et des positions relatives des graphes.

$$\sqrt{2-2x + x^2} - 1 \quad ; \quad \sin\big[\pi(x+1)\big] \quad ; \quad (x-1)\ln x \quad ; \quad e^{-1/(x-1)^2}$$

(poser $u = x-1$ et écrire un développement limité à l'ordre 4 des trois premières fonctions ; montrer que la dernière est infiniment petite devant toutes les puissances de $x-1$)

6. Calculons $I(x) = \int \sqrt{a^2 + t^2}\, dt$ en intégrant par parties (a est un nombre non nul).

a) Montrer que $I(x) = x\sqrt{a^2 + x^2} - J(x)$, avec $J(x) = \int \dfrac{t^2}{\sqrt{a^2+t^2}}\, dt$.

b) Montrer que $J(x) = \int \sqrt{a^2 + t^2}\, dt - \int \dfrac{a^2}{\sqrt{a^2+t^2}}\, dt$. En déduire la formule

$$\int \sqrt{a^2 + t^2}\, dt = \frac{1}{2} x\sqrt{a^2 + x^2} + \frac{a^2}{2} \ln\big[x + \sqrt{a^2 + x^2}\big].$$

@ **7. Enveloppe d'une famille de droites.** Dans un plan muni d'un repère, on se donne des droites D_t dépendant d'un paramètre réel t ; l'équation de D_t est de la forme $(D_t)\colon\ p(t)x + q(t)y + r(t) = 0$, où l'on suppose que les coefficients sont des fonctions

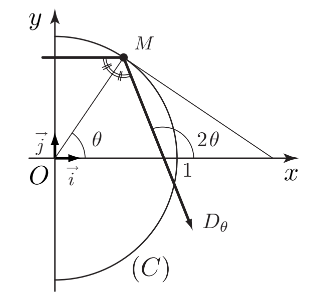
dérivables. *L'enveloppe des droites* D_t est une courbe paramétrée E qui, pour tout t, est tangente à D_t en un point $M_t = \big(x(t), y(t)\big)$.

a) Montrer qu'on a les égalités (1) : $p(t)x(t) + q(t)y(t) + r(t) = 0$ et
(2) : $p(t)x'(t) + q(t)y'(t) = 0$, pour tout t.

b) En dérivant (1), en déduire que $x(t)$ et $y(t)$ sont solutions du système linéaire

$$\begin{cases} p(t)x(t) + q(t)y(t) + r(t) & = 0 \\ p'(t)x(t) + q'(t)y(t) + r'(t) & = 0 \end{cases}$$

8. Un exemple d'enveloppe. Étant donné un repère orthonormé $(O;\vec{i},\vec{j})$ du plan, on considère le demi-cercle (C) de centre O et de rayon 1 situé du côté $x>0$. Imaginons un rayon lumineux parallèle à Ox venant frapper (C) en $M=(\cos\theta,\sin\theta)$: il se réfléchit dans une direction D_θ dépendant de θ.



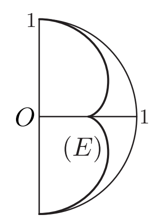
a) Montrer que si $0 \leqslant \theta \leqslant \pi/2$, la tangente au cercle en M fait l'angle $\pi/2 - \theta$ avec Ox ; en déduire que la droite D_θ fait l'angle 2θ avec Ox et que l'équation de D_θ est : $x \sin 2\theta - y \cos 2\theta - \sin\theta = 0$.

b) Montrer que lorsque θ varie de $-\pi/2$ à $\pi/2$, la droite D_θ enveloppe la courbe (E) : $\big(x(\theta), y(\theta)\big)$, où $x(\theta)$ et $y(\theta)$ sont solutions du système linéaire

$$\begin{cases} x\sin 2\theta - y\cos 2\theta - \sin\theta & = 0 \\ x\cos 2\theta + y\sin 2\theta - (1/2)\cos\theta & = 0 \end{cases}$$

En déduire que l'on a $x(\theta)=(3/2)\cos\theta-\cos^3\theta$ et $y(\theta)=\sin^3\theta$ pour $-\pi/2\leqslant\theta\leqslant\pi/2$.

c) Montrer que la courbe (E) est symétrique par rapport à Ox. En étudiant le sens de variation des fonctions $x(\theta)$ et $y(\theta)$, vérifier que (E) a l'allure ci-contre, où le point de rebroussement situé sur Ox est à l'abscisse $1/2$.



C'est la courbe qu'on peut voir briller au fond d'un récipient cylindrique éclairé de côté. Les courbes ainsi obtenues par réflexion sur des surfaces s'appellent des *caustiques*.

@ **9. Un cylindre évidé.** Le cylindre C d'axe Oz et de rayon 1 est formé des points à distance unité de l'axe : l'équation de C est donc $x^2 + y^2 = 1$. Évidons C de l'intérieur du cylindre C'_r de rayon r et d'axe Oy : l'équation de C'_r est $x^2 + z^2 = r^2$. Appelons γ_r la courbe d'intersection de C et de C'_r.

a) Les points de C ont pour coordonnées $(\cos t, \sin t, z)$: montrer que γ est formée des points $(\cos t, \sin t, \pm\sqrt{r^2 - \cos^2 t})$.

b) On suppose $r > 1$. Montrer que la courbe γ possède une tangente en tout point (figure **1**).

c) On suppose $0 < r < 1$. Montrer que γ est formée de deux courbes fermées disjointes symétriques par rapport au plan xOz et que chacune de ces courbes a une tangente en tout point (figure **2**).

d) Supposons que les deux cylindres ont le même rayon $r = 1$. Montrer qu'aux points $(\pm 1, 0, 0)$, la courbe γ se recoupe elle-même et trouver les vecteurs qui dirigent les tangentes en ces points ? (figure **3**).

On voit qu'il n'est pas possible d'usiner précisément un « té » formé de deux tubes ayant même diamètre et même épaisseur.

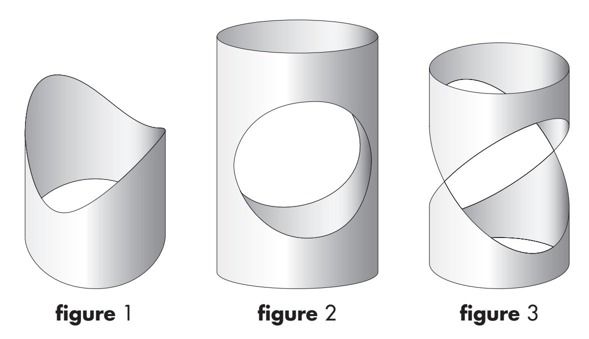


figure 1 **figure** 2 **figure** 3

@10. Au cours d'un trajet de longueur L, un promeneur perd successivement ses deux canifs, cela ayant pu se produire de façon équiprobable en n'importe quel endroit du parcours. On cherche la probabilité p pour qu'en refaisant le trajet, le promeneur retrouve l'un des canifs avant d'avoir effectué une distance d. Notons X_1 et X_2 la distance où se trouvent ces objets depuis le point de départ : X_1 et X_2 sont des variables aléatoires, de loi uniforme sur $[0, L]$.

a) Calculer la probabilité pour que $X_1 \geqslant d$ et la probabilité pour que X_1 et X_2 soient tous deux supérieurs ou égaux à d. En déduire $p = (d/L)\big[2-(d/L)\big]$.

b) Notons X la variable aléatoire : distance du point de départ au premier canif. Calculer la densité de X et montrer que l'espérance de X est $L/3$.

11. Calcul d'un intervalle de confiance. En fin de fabrication, on veut savoir quelle est la durée de vie des composants électroniques produits. Un contrôle de qualité montre que sur 200 composants, 5 ont une durée de vie inférieure à deux ans. On sait que la durée de vie X d'un composant suit une loi exponentielle : $P(X < t) = 1 - e^{-\lambda t}$, où λ est le nombre positif qu'on veut estimer.
Considérons la variable aléatoire Y qui vaut 1 si $X > 2$ et 0 si $X < 2$.

a) Montrer que l'espérance de Y est $E = e^{-2\lambda}$.

b) Représentons l'échantillon testé par la donnée $Y_1, \ldots, Y_n$, où $n = 200$. Montrer que la moyenne empirique $\overline{Y}_n = \dfrac{Y_1 + \cdots + Y_n}{n}$ vaut 195/200 et que l'écart-type de Y s'estime à $\sigma \simeq 0{,}156$. Montrer que E est compris entre $\overline{Y}_n - \dfrac{2\sigma}{\sqrt{200}}$ et $\overline{Y}_n + \dfrac{2\sigma}{\sqrt{200}}$ avec une probabilité supérieure à $0{,}95$. En déduire que λ est compris entre $0{,}001$ et $0{,}024$ avec un risque d'erreur inférieur à 5%.
En acceptant ce risque, à quelle proportion de composants le fabriquant peut-il garantir une durée de vie supérieure à cinq ans ?

c) Quelle durée de vie minimum peut-on raisonnablement garantir à 99% ?

@12. Des intégrales qui tendent vers 0 . Soit $f : [0,b] \to \mathbb{R}$ une fonction strictement croissante telle que $0 \leqslant f(t) \leqslant 1$ pour tout $t \in [0,b]$. On pose $J_n = \int_0^b \left[f(t)\right]^n dt$ pour tout entier $n \geqslant 1$.

a) Montrer que l'on a $\left[f(t)\right]^{n+1} \leqslant \left[f(t)\right]^n$ pour tout $n \geqslant 1$. En déduire que la suite (J_n) est décroissante et qu'elle a une limite $\ell \geqslant 0$.

b) Soit ε un nombre tel que $0 < \varepsilon < b$. Posons $a = b - \varepsilon$ et $r = f(a)$.

(i) Montrer que l'on a $0 \leqslant r < 1$ et $f(t) \leqslant r$ pour tout $t \in [0, a]$.

(ii) Montrer que pour tout $n \geqslant 1$, on a $0 \leqslant J_n \leqslant ar^n + \varepsilon$ (considérer l'intégrale de 0 à a et l'intégrale de a à b).

(iii) En passant à la limite quand n tend vers $+\infty$, en déduire $0 \leqslant \ell \leqslant \varepsilon$.

c) Montrer que $\lim J_n = 0$ (zéro est le seul nombre réel positif ou nul qui soit inférieur ou égal à tous les nombres $\epsilon > 0$).

13. Intégrales de Wallis. Pour tout entier $n \geqslant 0$, posons $I_n = \int_0^{\pi/2} (\sin t)^n \, dt$, où par convention $I_0 = \int_0^{\pi/2} 1 \, dt = \frac{\pi}{2}$.

a) En intégrant par parties $\int_0^{\pi/2} (\sin t)^n \sin t \, dt$, montrer que l'on a $I_{n+1} = n(I_{n-1} - I_{n+1})$ et en déduire $I_{n+1} = \frac{n}{n+1} I_{n-1}$ pour tout entier $n \geqslant 1$. Calculer I_1.

b) En déduire les formules suivantes :

$$I_{2p} = \frac{1.3.5 \cdots (2p-1)}{2.4.6 \cdots (2p)} \frac{\pi}{2} \quad \text{et} \quad I_{2p+1} = \frac{2.4.6 \cdots (2p)}{1.3.5 \cdots (2p+1)}.$$

c) Montrer que pour $t \in [0, \pi/2]$, on a $0 \leqslant (\sin t)^{n+1} \leqslant (\sin t)^n$ pour tout $n \geqslant 1$.

(i) En déduire que la suite (I_n) est décroissante et que l'on a $1 \geqslant \frac{I_{n+1}}{I_n} \geqslant \frac{I_{n+1}}{I_{n-1}} = \frac{n}{n+1}$.

(ii) Montrer que $\lim_{n \to +\infty} \frac{I_{n+1}}{I_n} = 1$.

d) En appliquant l'exercice précédent, montrer que $\lim_{n \to +\infty} I_n = 0$.

Chapitre 11

Interpolation, calcul numérique d'intégrales

1. Interpolation polynomiale

1.1 Les polynômes de Lagrange

Pour définir un polynôme $P = p_0 + p_1 x + \cdots + p_n x^n$ de degré n, il faut $n+1$ coefficients : on peut donc s'attendre à ce que P soit déterminé par $n+1$ équations. Donnons-nous $n+1$ nombres $x_0, x_1, \ldots, x_n$ (deux à deux différents) et des valeurs $y_0, y_1, \ldots, y_n$ quelconques, et cherchons un polynôme P dont le graphe passe par les points (x_i, y_i). Les égalités $P(x_i) = y_i$ s'écrivent sous forme d'un système de $n+1$ équations à $n+1$ inconnues :

$$\begin{cases} P(x_0) = p_0 + x_0 p_1 + x_0^2 p_2 + \cdots + x_0^n p_n & = y_0 \\ P(x_1) = p_0 + x_1 p_1 + x_1^2 p_2 + \cdots + x_1^n p_n & = y_1 \\ \qquad\qquad\vdots & \vdots \\ P(x_n) = p_0 + x_n p_1 + x_n^2 p_2 + \cdots + x_n^n p_n & = y_n \end{cases}$$

Puisque les inconnues sont les p_i, ce système d'équations est linéaire et l'on peut montrer que son déterminant est différent de 0. Par exemple, pour $n=2$, le déterminant est

$$\begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} = (x_0 - x_1)(x_1 - x_2)(x_2 - x_0) \neq 0$$

car nous avons supposé x_0, x_1, x_2 deux à deux différents.

Il y a une unique solution $(p_0, p_1, \ldots, p_n)$ et un unique polynôme P passant par les points (x_i, y_i). Nous allons montrer ce résultat et calculer le polynôme.

Cas du degré 1. Par les deux points $A_0=(x_0,y_0)$ et $A_1=(x_1,y_1)$, il passe une unique droite : sa pente est $\dfrac{y_1-y_0}{x_1-x_0}$, donc son équation est

$$y=y_0+\frac{y_1-y_0}{x_1-x_0}(x-x_0)=y_0\frac{x-x_1}{x_0-x_1}+y_1\frac{x-x_0}{x_1-x_0}$$

Le polynôme de degré 1 qui passe par A_0 et A_1 est donc $P=y_0\dfrac{x-x_1}{x_0-x_1}+y_1\dfrac{x-x_0}{x_1-x_0}$.

Cas $n=2$. Cherchons le polynôme sous la forme

$$P=a_0(x-x_1)(x-x_2)+a_1(x-x_0)(x-x_2)+a_2(x-x_0)(x-x_1)$$

On a $P(x_0)=a_0(x_0-x_1)(x_0-x_2)$, donc l'égalité $P(x_0)=y_0$ est réalisée si l'on prend $a_0=\dfrac{y_0}{(x_0-x_1)(x_0-x_2)}$. En posant de même $a_1=\dfrac{y_1}{(x_1-x_0)(x_1-x_2)}$ et $a_2=\dfrac{y_2}{(x_2-x_0)(x_2-x_1)}$, on obtient la solution

$$P=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

En général, P est de degré 2, donc son graphe est une parabole. Mais si les points sont alignés, alors P est de degré 1 et son graphe est une droite.

Cas général. On définit le *polynôme de Lagrange* L_k qui prend la valeur 1 en x_k et s'annule en tous les x_i tels que $i\neq k$:

$$L_k(x)=\prod_{i\neq k}\frac{x-x_i}{x_k-x_i}$$

Le symbole $\prod_{i\neq k}$ signifie que l'on fait le produit des n facteurs pour i tel que $0\leqslant i\leqslant n$ et $i\neq k$. Chaque facteur étant de degré 1, le polynôme L_k est de degré n. Pour $x=x_k$, le numérateur et le dénominateur sont identiques, donc $L_k(x_k)=1$. Pour $i\neq k$ et $x=x_i$, on a $L_k(x_i)=0$ car $(x-x_i)$ est un facteur au numérateur.

Définissons le polynôme P en posant

$$P=y_0L_0+y_1L_1+\cdots+y_nL_n\,.$$

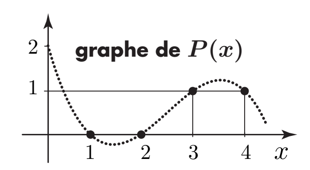
On a alors $P(x_0)=y_0L_0(x_0)+y_1L_1(x_0)+\cdots+y_nL_n(x_0)=y_0\times1+y_1\times0+\cdots+y_n\times0=y_0$, et de même $P(x_k)=y_k$ pour tout k. Ce polynôme répond donc à la question.

Définition

Le polynôme P s'appelle le *polynôme d'interpolation* pour les points (x_i,y_i).

Supposons que Q est un polynôme de degré inférieur ou égal à n tel que $Q(x_i)=y_i$ pour tout $i=0,1,\dots,n$. Si $P-Q$ n'est pas le polynôme nul, son degré est au plus n, donc $P-Q$ possède au plus n racines (page 46). Or le polynôme $P-Q$ s'annule en les $n+1$ valeurs $x_0,\dots,x_n$: c'est donc que $P-Q$ est le polynôme nul, autrement dit $P=Q$. Cela montre que P est le seul polynôme de degré au plus n vérifiant $P(x_i)=y_i$ quel que soit i.

Exemple. Le polynôme d'interpolation pour les points $(x_0,y_0)=(1,0)$, $(x_1,y_1)=(2,0)$, $(x_2,y_2)=(3,1)$, $(x_3,y_3)=(4,1)$ est $P = \frac{1}{6}\left(-2x^3 + 15x^2 - 31x + 18\right)$.



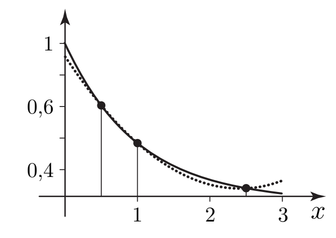
Interpolation d'une fonction

Soit f une fonction. Si l'on connaît les valeurs $y_i = f(x_i)$ en $n+1$ points $x_0, x_1, \ldots, x_n$, le polynôme d'interpolation pour les données (x_i, y_i) prend les mêmes valeurs que f en $x_0, x_1, \ldots, x_n$.

On dit que P est le *polynôme d'interpolation de* f *en* $x_0, x_1, \ldots, x_n$.

Si pour les autres valeurs de x l'écart entre $P(x)$ et $f(x)$ n'est pas trop grand, on peut prendre $P(x)$ comme valeur approchée de $f(x)$, du moins quand x reste dans un segment contenant les x_i.

Le calcul d'une valeur $P(x)$ ne demande que des multiplications et des additions : il est donc souvent plus rapide que celui de $f(x)$, surtout lorsque la fonction f a une expression compliquée ou faisant intervenir des exponentielles ou des fonctions trigonométriques. De plus, on peut optimiser le calcul de $P(x)$ en employant la méthode de Hörner (page 48).



La figure ci-contre montre le graphe de la fonction e^{-x} et son polynôme d'interpolation (de degré 2) en $x_0 = 0{,}5$, $x_1 = 1$, $x_2 = 2{,}5$.

Cas d'une fonction f connue par des valeurs discrètes. Il arrive que la fonction f ne soit connue que par les valeurs $y_0, y_1, \ldots, y_n$ qu'elle prend en certains points $x_0, x_1, \ldots, x_n$. On ne dispose alors d'aucune formule générale permettant de calculer d'autres valeurs de la fonction. Dans ce cas, $P(x)$ peut constituer une formule approximative, mais raisonnable, pour la valeur $f(x)$, à condition que x reste dans un intervalle convenable.

Calcul du polynôme d'interpolation

Voici une méthode pour calculer le polynôme d'interpolation relatif à des points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$.

Notation. Soient $m_1, m_2, \ldots, m_k$ des entiers distincts compris entre 0 et n. On note $P_{m_1,m_2,\ldots,m_k}$ le polynôme d'interpolation aux k points d'abscisses $x_{m_1}, x_{m_2}, \ldots, x_{m_k}$.

Une formule commode. Soient x_i et x_j deux nombres distincts parmi $\{x_0,\ldots,x_n\}$. Le polynôme d'interpolation en $x_0, x_1, \ldots, x_k$ est alors

$$(*) \qquad P = \frac{1}{x_i - x_j}\Big[(x - x_j)P_{0,\ldots,j-1,j+1,\ldots,k} - (x - x_i)P_{0,\ldots,i-1,i+1,\ldots,k}\Big]$$

Posons pour simplifier $U = P_{0,\ldots,j-1,j+1,\ldots,k}$ et $V = P_{0,\ldots,i-1,i+1,\ldots,k}$. Puisque U et V sont des polynômes d'interpolation en k points, ils sont de degré au plus $k-1$, donc P est de

degré au plus k. Soit q un entier tel que $0 \leqslant q \leqslant k$, $q \neq i$ et $q \neq j$. On a $U(x_q) = y_q = V(x_q)$, donc

$$P(x_q) = \frac{1}{x_i - x_j}\big[(x_q - x_j)U(x_q) - (x_q - x_i)V(x_q)\big] = \frac{1}{x_i - x_j}(x_q - x_j - x_q + x_i)y_q = y_q$$

On a aussi

$$P(x_i) = \frac{1}{x_i - x_j}\big[(x_i - x_j)U(x_i) - (x_i - x_i)V(x_i)\big] = U(x_i) = y_i$$

et de même $P(x_j) = y_j$.

Méthode. Supposons qu'on a calculé le polynôme $P_{0,1}$ (interpolation en x_0, x_1) et le polynôme $P_{1,2}$ (interpolation en x_1, x_2). D'après la formule ci-dessus, le polynôme d'interpolation en x_0, x_1, x_2 est $P_{0,1,2} = \frac{1}{x_2 - x_0}\big[(x - x_0)P_{1,2} - (x - x_2)P_{0,1}\big]$.

Afin d'éviter des indices trop compliqués, posons $Q_{i,0} = y_i$ et

$Q_{i,d}$ le polynôme d'interpolation aux $d+1$ points $x_{i-d}, x_{i-d+1}, \dots, x_i$, si $i \geqslant d \geqslant 1$.

Supposons qu'on ait calculé tous les polynômes $Q_{i,1}$ pour $i \geqslant 1$, c'est-à-dire tous les polynômes d'interpolation en deux points consécutifs x_{i-1}, x_i. Alors d'après la formule, le polynôme d'interpolation en trois points consécutifs x_{i-2}, x_{i-1}, x_i est

$$Q_{i,2} = \frac{1}{x_i - x_{i-2}}\big[(x - x_{i-2})Q_{i,1} - (x - x_i)Q_{i-1,1}\big]$$

(on a appliqué la formule $(*)$ en choisissant les points extrêmes x_{i-2} et x_i parmi les trois points x_{i-2}, x_{i-1}, x_i.)

Le polynôme $Q_{i,3}$ d'interpolation en quatre points consécutifs $x_{i-3}, x_{i-2}, x_{i-1}, x_i$ s'exprime maintenant au moyen des polynômes d'interpolation en x_{i-2}, x_{i-1}, x_i et en $x_{i-3}, x_{i-2}, x_{i-1}$, c'est-à-dire au moyen de $Q_{i,2}$ et $Q_{i-1,2}$. De manière générale, si l'on connaît tous les polynômes $Q_{i,d-1}$, alors

$$Q_{i,d} = \frac{1}{x_i - x_{i-d}}\big[(x - x_{i-d})Q_{i,d-1} - (x - x_i)Q_{i-1,d-1}\big]$$

Avec notre définition $Q_{i,0} = y_i$, cette égalité est encore vraie pour $d = 1$. On peut ainsi calculer de proche en proche $Q_{n,n}$ qui est le polynôme d'interpolation aux points $x_0, x_1, \dots, x_n$, c'est-à-dire le polynôme cherché.

Algorithme de Neuville

initialisation :

variable x

nombres $x_0, x_1, \dots, x_n$ deux à deux différents

nombres $y_0, y_1, \dots, y_n$

$Q_{i,0} \leftarrow y_i$ pour $0 \leqslant i \leqslant n$

boucle : pour $i = 1, 2, \dots n$, faire

pour $d = 1, 2 \dots, i$: $Q_{i,d} \leftarrow \frac{1}{x_i - x_{i-d}}\big[(x - x_{i-d})Q_{i,d-1} - (x - x_i)Q_{i-1,d-1}\big]$

fin : $Q_{n,n}$ est le polynôme d'interpolation pour les points (x_i, y_i).

On pratique souvent l'algorithme en prenant pour x un nombre fixé : on obtient alors la valeur en x du polynôme d'interpolation.

Cet algorithme présente un avantage précieux : au cours du calcul de $Q_{n,n}$, on a obtenu tous les polynômes $Q_{n,d}$ pour $d \leqslant n$; on peut donc ajouter un point supplémentaire et poursuivre l'algorithme en utilisant les polynômes déjà calculés.

Supposons par exemple que pour une fonction f, on veuille calculer une valeur inconnue $f(a)$ en interpolant à partir de valeurs connues $f(x_i)=y_i$. Si l'on s'est fixé une précision ε, on peut ajouter des points tant que $|Q_{n,n}(a) - Q_{n-1,n-1}(a)|$ dépasse ε.

Erreur d'interpolation

Lorsqu'on interpole une fonction, il est bon de connaître une majoration de l'erreur $|f(x)-P(x)|$, où P est le polynôme d'interpolation de f en $x_0,x_1,\ldots,x_n$. Supposons que f possède une dérivée $(n+1)$-ième continue sur un intervalle $[a,b]$ contenant les x_i.

Pour tout $x \in [a,b]$, *on a* $f(x) - P(x) = \dfrac{f^{(n+1)}(c)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$, *où* c *est un nombre dépendant de* x *et compris entre* a *et* b.

Si l'on connaît un nombre M tel que $\left|f^{(n+1)}(t)\right| \leqslant M$ pour tout $t \in [a,b]$, alors l'erreur d'interpolation en tout point x de $[a,b]$ est au plus égale à

$$\max_{x\in[a,b]} \left|f(x) - P(x)\right| \leqslant \frac{M}{(n+1)!} \max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i|$$

Ce résultat n'a qu'un intérêt théorique, car même dans le cas où f est définie par une formule, il est rare qu'on fasse une estimation de la dérivée $(n+1)$-ième.

Démonstration. Si x est l'un des x_i, les deux membres de l'égalité sont nuls. Supposons x différent de tous les x_i, posons $Q(t) = \prod_{i=0}^{n} (t - x_i)$ et

$$g(t) = f(t) - P(t) - Q(t)\frac{f(x) - P(x)}{Q(x)} \quad \text{pour tout } t \in [a,b].$$

On a $g(x) = 0$, $Q(x_i) = 0$ et $f(x_i) = P(x_i)$, donc la fonction g s'annule au moins en les $n+2$ points $x, x_0, \ldots, x_n$. D'après le théorème des accroissements finis, la dérivée g' s'annule dans chacun des $n+1$ intervalles délimités par ces points, donc g' s'annule au moins $n+1$ fois. En répétant ce raisonnement, on en conclut que la dérivée $g^{(n+1)}$ s'annule en au moins un point c de $[a,b]$. La fonction polynôme $t \mapsto P(t)$ étant de degré n, sa dérivée $(n+1)$-ième est nulle, donc

$$g^{(n+1)}(t) = f^{(n+1)}(t) - Q^{(n+1)}(t)\frac{f(x) - P(x)}{Q(x)} = f^{(n+1)}(t) - (n+1)!\frac{f(x) - P(x)}{Q(x)}$$

On a ainsi $0 = f^{(n+1)}(c) - (n+1)!\dfrac{f(x) - P(x)}{Q(x)}$ et le résultat. ■

1.2 Choix des points d'interpolation

Lorsqu'on veut approximer une fonction f sur un segment $[a,b]$, il est naturel de choisir les points d'interpolation $x_0, x_1, \ldots, x_n$ équirépartis dans le segment, c'est-à-dire de poser $h=(b-a)/n$ et $x_0=a$, $x_1=a+h, \ldots, x_i=a+ih, \ldots$, $x_n=b$. Mais pour un nombre de points donné, ce n'est pas ainsi qu'on obtient, en général, la meilleure précision.

Les polynômes de Chebychev

Pour tout $x \in [-1,1]$, posons $\cos\theta = x$, où $0 \leqslant \theta \leqslant \pi$. Si n est un entier positif, posons $T_n(x) = \cos(n\theta)$. Ainsi $T_0(x) = 1$ et $T_1(x) = x$. En utilisant la formule d'addition pour $\cos(a+b)$, on a $\cos\big((n+1)\theta\big) + \cos\big((n-1)\theta\big) = 2\cos\theta\cos(n\theta)$ ou encore

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

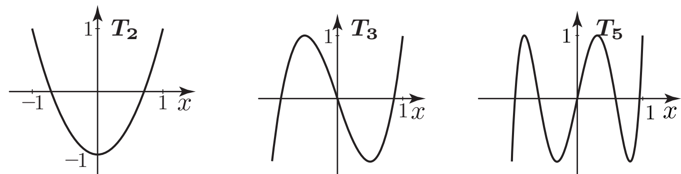
Cette relation permet de calculer $T_n(x)$ de proche en proche à partir de $T_0(x)$ et de $T_1(x)$. Par exemple,

$$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1 \quad \text{et} \quad T_3(x) = 2xT_2(x) - T_1(x) = 4x^3 - 3x$$

La formule montre aussi que T_n est un polynôme de degré n et que son coefficient dominant est 2^{n-1} (raisonnement par récurrence).

Les polynômes T_n s'appellent les polynômes de Chebychev.

Les racines de T_n sont les nombres $x = \cos\theta$ tels que $\cos(n\theta) = 0$, c'est-à-dire $n\theta = \dfrac{(2k-1)\pi}{2}$, où $k = 1, 2, \ldots, n$. Il y a donc n racines réelles $r_k = \cos\left(\dfrac{2k-1}{2n}\pi\right)$, où $k = 1, 2, \ldots, n$.



Cherchons les extrema de T_n sur le segment $[-1,1]$. On a

$$\frac{dT_n}{d\theta} = \frac{dT_n}{dx}\frac{dx}{d\theta} = \frac{dT_n}{dx}\frac{d(\cos\theta)}{d\theta} = \frac{dT_n}{dx}(-\sin\theta)$$

Pour $x \in\,]-1,1[$, $\sin\theta$ n'est pas nul, donc

$$\frac{dT_n}{dx} = 0 \quad \text{si et seulement si} \quad \frac{dT_n}{d\theta} = \frac{d(\cos n\theta)}{d\theta} = -n\sin(n\theta) = 0\,,$$

ce qui donne $\theta = \dfrac{k\pi}{n}$, où $k = 1, 2, \ldots, n-1$. Aux points $z_k = \cos(k\pi/n)$, l'extremum vaut $T_n(z_k) = \cos\big(n(k\pi/n)\big) = \cos(k\pi) = (-1)^k$. Puisque $T_n(1) = 1$ (pour $\theta = 0$) et $T_n(-1) = (-1)^n$ (pour $\theta = \pi$), on en déduit la propriété suivante.

Sur $[-1,1]$, $T_n(x)$ atteint ses extrema aux points $z_k = \cos(k\pi/n)$, où $k = 0, 1, \ldots, n$, et $T_n(z_k) = (-1)^k$.

Voici une propriété des polynômes de Chebychev. Posons $\tilde{T}_n = \dfrac{1}{2^{n-1}}T_n$. Le coefficient dominant du polynôme $\tilde{T}_n$ est égal à 1 : c'est un polynôme *unitaire* et de degré n.

Propriété. *Pour tout polynôme P unitaire et de degré $n \geqslant 1$, on a*

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} \left|\tilde{T}_n(x)\right| \leqslant \max_{x\in[-1,1]} \left|P_n(x)\right|$$

De plus, on n'a l'égalité que si $P = \tilde{T}_n$.

Démonstration. Supposons que P est un polynôme unitaire de degré n et que

$$\max_{x\in[-1,1]} \left|P_n(x)\right| \leqslant \frac{1}{2^{n-1}} = \max_{x\in[-1,1]} \left|\tilde{T}_n(x)\right|.$$

Le polynôme $Q=\tilde{T}_n - P$ est de degré au plus $n-1$, car le terme x^n disparaît dans la différence. On a $Q(z_k) = \tilde{T}_n(z_k) - P(z_k) = \frac{(-1)^k}{2^{n-1}} - P(z_k)$. Par hypothèse, on a $\left|P(z_k)\right| \leqslant \frac{1}{2^{n-1}}$, donc $Q(z_k) \leqslant 0$ si k est impair et $Q(z_k) \geqslant 0$ si k est pair. Par le théorème des valeurs intermédiaires, on en déduit que Q s'annule au moins une fois entre z_k et z_{k+1}, pour $k = 0, 1, \ldots, n-1$, donc possède au moins n racines dans $[-1, 1]$. Comme un polynôme de degré $n-1$ a au plus $n-1$ racines, cela n'est possible que si $Q = 0$, autrement dit $P = \tilde{T}_n$. ■

Application à l'erreur d'interpolation

- Supposons que f est une fonction définie sur $[-1, 1]$. Quand on interpole f aux points $x_0, x_1, \ldots, x_n$ de $[-1, 1]$, l'erreur provenant du choix des points x_i réside dans le facteur $\max_{x\in[-1,1]} \left|\Pi(x)\right|$, où $\Pi(x) = \prod_{i=0}^{n} (x - x_i)$ est un polynôme unitaire de degré $n+1$. D'après la propriété précédente, ce terme sera minimum si $\Pi = \tilde{T}_{n+1} = (x - r_1)(x - r_2) \cdots (x - r_{n+1})$, c'est-à-dire si l'on choisit

$$x_k = r_{k+1} = \cos\left(\frac{2k+1}{2(n+1)}\pi\right), \quad \text{pour } k = 0, 1, \ldots, n.$$

  Puisque le maximum de $|\tilde{T}_{n+1}| = \frac{1}{2^n}|T_{n+1}|$ sur $[-1, 1]$ est $\frac{1}{2^n}$, on en déduit que si P est le polynôme d'interpolation utilisant ces $n+1$ points, on a la majoration

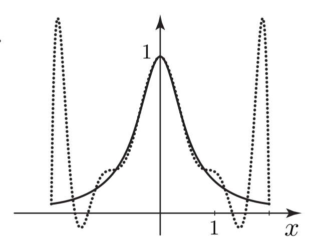
$$\max_{x\in[-1,1]} \left|f(x) - P(x)\right| \leqslant \frac{1}{2^n(n+1)!} \max_{x\in[-1,1]} \left|f^{(n+1)}(x)\right|$$

- Pour une fonction définie sur un segment $[a, b]$ quelconque, on transporte ces abscisses par l'application affine $u : [-1, 1] \to [a, b]$ qui envoie -1 sur a et 1 sur b ; elle est définie par $u(x) = \frac{1}{2}\left[(b - a)x + a + b\right]$.

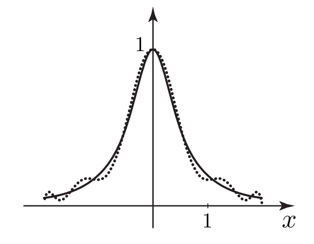
  Règle : *Pour une meilleure interpolation, choisir les $n+1$ points d'abscisses $x'_k = u(x_k)$.*

Exemple. Soit la fonction $f(x) = \frac{1}{1 + 4x^2}$ entre -2 et 2.

- La figure ci-contre montre le graphe de f et celui du polynôme d'interpolation en les onze points $x_i = (0,4)i$, où $-5 \leqslant i \leqslant 5$. Aux bords du segment, l'approximation est très mauvaise.



► Interpolons maintenant la fonction en choisissant les abscisses de Chebychev sur $[-2, 2]$: elles sont définies par $x'_k = 2\cos\dfrac{2k-1}{11}\dfrac{\pi}{2}$, où $1 \leqslant k \leqslant 11$. On voit que, pour le même nombre de points, l'approximation est sensiblement meilleure vers les bords du segment.



1.3 Interpolation par des fonctions splines

Quand on interpole une fonction f en utilisant un polynôme P de grand degré n, le graphe de P présente souvent des oscillations qui ne reflètent pas du tout l'allure de la fonction. Pour remédier à cet inconvénient, nous allons interpoler avec des polynômes de degré 3 en tenant compte de la dérivée de f.

Interpolation avec trois points

Donnons-nous trois points $x_0 < x_1 < x_2$ dans l'intervalle de définition de f. Nous allons construire une fonction S par morceaux sur chacun des intervalles $[x_0, x_1]$, $[x_1, x_2]$ ayant les propriétés suivantes :

A) sur $[x_0, x_1]$, $S = S_0$ est une fonction polynôme de degré au plus 3 ; de même sur $[x_1, x_2]$, $S = S_1$ est polynomiale de degré au plus 3.

B) pour $i=0,1,2$, $S(x_i)=f(x_i)$, donc en particulier $S_0(x_1)=f(x_1)=S_1(x_1)$ (conditions d'interpolation)

C) $S'_0(x_1) = S'_1(x_1)$ (raccordement des tangentes)

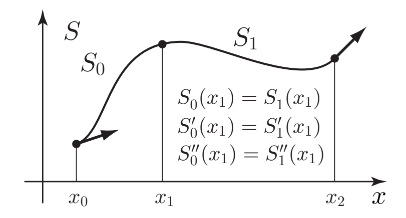
D) $S''_0(x_1) = S''_1(x_1)$ (raccordement des courbures)

E) l'une des conditions suivantes est vérifiée :

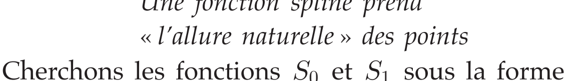
i) $S''(x_0) = S''(x_2) = 0$ (courbure nulle aux extrémités)

ii) $S'(x_0) = f'(x_0)$ et $S'(x_2) = f'(x_2)$ (tangente imposée aux extrémités).

Une telle fonction S s'appelle une *fonction spline cubique* pour ces données. Si l'on impose la condition (E) (i), on dit que la fonction spline est *naturelle* ; si l'on impose la condition (E) (ii), on dit que la fonction spline est *contrainte*.



Une fonction spline contrainte


Une fonction spline prend « l'allure naturelle » des points

Cherchons les fonctions S_0 et S_1 sous la forme

$$S_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3 \quad \text{pour } x \in [x_0, x_1]$$

$$S_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3 \quad \text{pour } x \in [x_1, x_2]$$

Les conditions précédentes s'expriment par des équations linéaires aux huit inconnues a_i, b_i, c_i, d_i, $i = 1$ ou 2. Il y a aussi huit équations : quatre pour l'interpolation, une pour le raccord des dérivées premières en x_1, une pour le raccord des dérivées secondes et deux pour la condition aux extrémités.

Les nombres a_0 et a_1 sont déterminés par $a_0 = S_0(x_0) = f(x_0)$ et $a_1 = S_1(x_1) = f(x_1)$. Posons $h_0 = x_1 - x_0$, $h_1 = x_2 - x_1$, $a_2 = f(x_2)$ et $c_2 = (1/2)S_1''(x_2)$.
Puisque $S_0''(x) = 2c_0 + 6d_0(x - x_0)$ et $S_1''(x) = 2c_1 + 6d_1(x - x_1)$, il vient $S_0''(x_0) = 2c_0$, $S_0''(x_1) = 2c_0 + 6d_0h_0$, $S_1''(x_1) = 2c_1$, $S_1''(x_2) = 2c_1 + 6d_1h_1$. La condition (D) et la définition de c_2 donnent alors les deux équations :

$$(1) \quad c_1 = c_0 + 3d_0h_0 \qquad , \qquad (2) \quad c_2 = c_1 + 3d_1h_1$$

Les conditions d'interpolation $a_1 = f(x_1) = S_0(x_1)$ et $a_2 = f(x_2) = S_1(x_2)$ s'écrivent

$$(3) \quad a_1 - a_0 = b_0h_0 + c_0h_0^2 + d_0h_0^3 \qquad , \qquad (4) \quad a_2 - a_1 = b_1h_1 + c_1h_1^2 + d_1h_1^3$$

Puisque $S_1'(x_1) = b_1$ et $S_0'(x_1) = b_0 + 2c_0h_0 + 3d_0h_0^2$, la condition (C) s'exprime par

$$(5) \quad b_1 - b_0 = 2c_0h_0 + 3d_0h_0^3$$

Tirons d_0 et d_1 de (1) et (2) et reportons dans (3), (4) et (5) :

$$(3') \quad a_1 - a_0 = b_0h_0 + \frac{h_0^2}{3}(2c_0 + c_1) \qquad , \qquad (4') \quad a_2 - a_1 = b_1h_1 + \frac{h_1^2}{3}(2c_1 + c_2)$$

$$(5') \quad b_1 - b_0 = h_0(c_0 + c_1)$$

En tirant b_0 et b_1 de $(3')$ et $(4')$ et en portant dans $(5')$, on obtient

$$b_1 - b_0 = \frac{a_2 - a_1}{h_1} - \frac{a_1 - a_0}{h_0} - \frac{h_1}{3}(2c_1 + c_2) + \frac{h_0}{3}(2c_0 + c_1) = h_0(c_0 + c_1)$$

c'est-à-dire

$$(*) \qquad h_0c_0 + 2(h_0 + h_1)c_1 + h_1c_2 = \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0)$$

► Si l'on choisit la condition (E) (i), alors $2c_0 = S_0''(x_0) = 0$, $2c_2 = S_1''(x_2) = 0$ et les coefficients c_0, c_1, c_2 sont solutions du système linéaire

$$\begin{bmatrix} 1 & 0 & 0 \\ h_0 & 2(h_0+h_1) & h_1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ (3/h_1)(a_2-a_1)-(3/h_0)(a_1-a_0) \\ 0 \end{bmatrix}$$

Le second membre est connu, car $a_0 = f(x_0)$, $a_1 = f(x_1)$ et $a_2 = f(x_2)$. On détermine ainsi c_1, puis les coefficients d_0 et d_1 au moyen de (1) et (2), et enfin b_0 et b_1 au moyen de $(3')$ et $(4')$.

► Si l'on a choisi la condition (E) (ii), alors on a $f'(x_0) = b_0 = \dfrac{a_1 - a_0}{h_0} - \dfrac{h_0}{3}(2c_0 + c_1)$ en utilisant $(3')$, c'est-à-dire

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(x_0)$$

De même, on a $f'(x_2) = S_1'(x_2) = b_1 + 2c_1h_1 + 3d_1h_1^2 = b_1 + 2c_1h_1 + h_1(c_2 - c_1) = b_1 + h_1(c_1 + c_2)$ en utilisant (2), et il vient

$$h_1c_1 + 2h_1c_2 = 3f'(x_2) - \frac{3}{h_1}(a_2 - a_1)$$

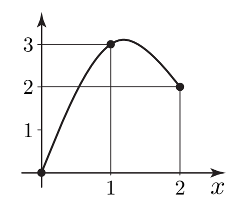
En tenant compte de $(*)$, les coefficients c_0, c_1, c_2 sont solutions du système linéaire

$$\begin{bmatrix} 2h_0 & h_0 & 0 \\ h_0 & 2(h_0+h_1) & h_1 \\ 0 & h_1 & 2h_1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} (3/h_0)(a_1-a_0)-3f'(x_1) \\ (3/h_1)(a_2-a_1)-(3/h_0)(a_1-a_0) \\ 3f'(x_2)-(3/h_1)(a_2-a_1) \end{bmatrix}$$

La matrice du système est à diagonale strictement dominante, donc est inversible (page 250) : il y a une unique solution. Après avoir résolu ce système, on en déduit d_0, d_1, b_0, b_1 comme précédemment.

Exemple. La fonction spline naturelle passant par les points $(0,0)$, $(1,3)$ et $(2,2)$ est définie par

$$S(x) = \begin{cases} 4x - x^3 & \text{si } 0 \leqslant x \leqslant 1 \\ x^3 - 6x^2 + 10x - 2 & \text{si } 1 \leqslant x \leqslant 2 \end{cases}$$



Cas général

Supposons que la fonction f est définie sur $[a,b]$ et donnons-nous des nombres

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b, \quad \text{avec } n \geqslant 2$$

Une *fonction spline cubique interpolante* pour ces données est une fonction S définie sur $[a,b]$ ayant les propriétés suivantes :

A) sur chaque intervalle $[x_i, x_{i+1}]$, où $i = 0, 1, \ldots, n-1$, $S = S_i$ est une fonction polynôme de degré au plus 3

B) $S(x_i) = f(x_i)$ pour $i = 0, 1, \ldots, n$ (donc en particulier $S_i(x_{i+1}) = S_{i+1}(x_{i+1})$)

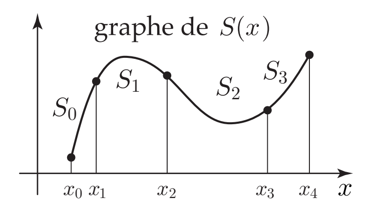
C) $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$ pour $i = 0, 1, \ldots, n-2$ (raccordement des tangentes)

D) $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$ pour $i = 0, 1, \ldots, n-2$ (raccordement des courbures)

E) l'une des conditions suivantes est vérifiée :

  i) $S''(a) = S''(b) = 0$ (tangente libre aux bords : spline naturelle)

  ii) $S'(a) = f'(a)$ et $S'(b) = f'(b)$ (tangente imposée aux bords : spline contrainte).



On cherche les fonctions $S_0, S_1, \cdots, S_{n-1}$ sous la forme

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

où $x \in [x_i, x_{i+1}]$ et $0 \leqslant i \leqslant n-1$.

- Les coefficients a_i sont donnés par $a_i = S_i(x_i) = f(x_i)$, pour $0 \leqslant i \leqslant n-1$.
- On pose $a_n = f(x_n) = f(b)$ et $h_i = x_{i+1} - x_i$ pour $i = 0, 1, \ldots, n-1$.
- Pour calculer $c_0, c_1, \cdots, c_{n-1}$, on résout un système linéaire $Ac = u$, où $c = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}$:
  - avec l'option (E) (i),

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \cdots & \vdots \\ 0 & h_1 & 2(h_1+h_2) & h_2 & & \vdots \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}$$

$$u = \begin{bmatrix} 0 \\ (3/h_1)(a_2-a_1)-(3/h_0)(a_1-a_0) \\ \vdots \\ (3/h_{n-1})(a_n-a_{n-1})-(3/h_{n-2})(a_{n-1}-a_{n-2}) \\ 0 \end{bmatrix}$$

• avec l'option (E) (ii),

$$A = \begin{bmatrix} 2h_0 & h_0 & 0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \cdots & \vdots \\ 0 & h_1 & 2(h_1+h_2) & h_2 & & \vdots \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix}$$

$$u = \begin{bmatrix} (3/h_0)(a_1-a_0)-3f'(a) \\ (3/h_1)(a_2-a_1)-(3/h_0)(a_1-a_0) \\ \vdots \\ (3/h_{n-1})(a_n-a_{n-1})-(3/h_{n-2})(a_{n-1}-a_{n-2}) \\ 3f'(b)-(3/h_{n-1})(a_n-a_{n-1}) \end{bmatrix}$$

Ces matrices A étant à diagonale strictement dominante, le système a une unique solution que l'on peut calculer par une méthode de relaxation (chapitre 8).

► Les coefficients b_i et d_i se calculent au moyen des a_i et des c_i par les formules analogues à (3′), (4′), (1) et (2) :

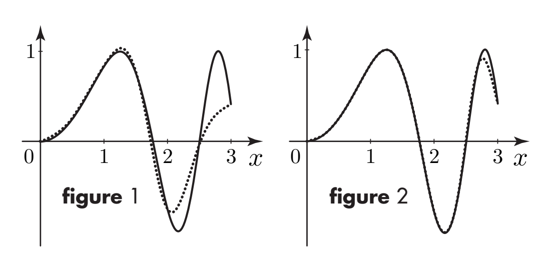
$$b_i = \frac{1}{h_i}(a_{i+1} - a_i) - \frac{h_i}{3}(2c_i + c_{i+1}) \quad \text{pour } 0 \leqslant i \leqslant n-1$$
$$d_i = \frac{1}{3h_i}(c_{i+1} - c_i) \quad \text{pour } 0 \leqslant i \leqslant n-1.$$

► Quand on cherche une fonction spline naturelle, on a seulement besoin de connaître les coordonnées des points d'interpolation : on peut donc faire le calcul lorsque la fonction n'est connue que par les valeurs qu'elle prend en un nombre fini de points $x_0, \ldots, x_n$. On obtient alors une formule analytique acceptable et valable partout dans $[x_0, x_n]$. Cela s'applique notamment lorsque la courbe est donnée par son seul dessin : en quadrillant le plan et en choisissant quelques points caractéristiques, on en déduit une bonne approximation analytique.

► Il faut peu de données pour décrire une fonction spline : seulement quatre coefficients numériques pour chacun des polynômes de degré 3 qui la constituent. Le codage de la fonction est donc particulièrement économique. Dans une imprimante, chaque lettre est ainsi mise en mémoire sous la forme des coefficients des fonctions splines qui en décrivent le contour.

Exemple. Prenons la fonction $f(x) = \sin(x^2)$ sur le segment $[0,3]$.

- Avec les sept points d'abscisse $x_i = (0{,}5)i$, pour $i = 0, 1, \ldots, 6$, on obtient, par une courbe spline naturelle, l'approximation montrée figure 1.
- En prenant les onze points d'abscisse $x_i = (0{,}3)i$, pour $i = 0, 1, \ldots, 10$, on obtient la figure 2.



Approximation par une courbe paramétrée

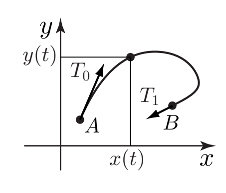
Quand la courbe n'est pas le graphe d'une fonction $y(x)$, on ne peut pas l'approcher par une fonction spline. Dans ce cas, on cherche comme approximation une courbe paramétrée $\big(x(t), y(t)\big)$, où $x(t)$ et $y(t)$ sont des polynômes de degré 3.

Donnons-nous :

- deux points A et B, de coordonnées (x_0, y_0) et (x_1, y_1)
- et des vecteurs $T_0 = (u_0, v_0)$ et $T_1 = (u_1, v_1)$: ce seront les vecteurs tangents à la courbe en A et B.

Pour $0 \leqslant t \leqslant 1$, posons

$$x(t)=\big[2(x_0-x_1)+(u_0+u_1)\big]t^3+\big[3(x_1-x_0)-(u_1+2u_0)\big]t^2+u_0t+x_0$$
$$y(t)=\big[2(y_0-y_1)+(v_0+v_1)\big]t^3+\big[3(y_1-y_0)-(v_1+2v_0)\big]t^2+v_0t+y_0$$



Cette courbe paramétrée a pour origine le point A (en $t = 0$), pour extrémité le point B (en $t = 1$), son vecteur tangent en A est T_0 et son vecteur tangent en B est T_1.

2. Calcul numérique d'intégrales

Dans la plupart des intégrales qu'on rencontre, il n'est pas possible d'exprimer une primitive à l'aide des fonctions usuelles : on a alors recours au calcul numérique.

Pour estimer la valeur d'une intégrale $\int_a^b f(t)\,dt$, on approche la fonction f par une fonction φ dont l'intégrale est très simple à calculer. Puisque $\left|\int_a^b f(t)\,dt - \int_a^b \varphi(t)\,dt\right| \leqslant \int_a^b |f(t) - \varphi(t)|\,dt$, il faut que l'intégrale $\int_a^b |f(t)-\varphi(t)|\,dt$ soit petite : la fonction φ peut différer notablement de f, mais sur de petits intervalles.

On peut choisir pour φ une fonction en escalier, comme celles qui permettent, par passage à la limite, de définir l'intégrale (page 286). On fait ainsi apparaître des rectangles dont la somme des surfaces est une valeur approchée de l'intégrale d'autant plus précise que leur base est plus petite.

Une formule d'intégration exacte

Pour toute fonction polynôme P de degré au plus 3, on a

$$\int_a^b P(t)\,dt = \frac{b-a}{6}\left[P(a) + 4P\left(\frac{a+b}{2}\right) + P(b)\right]$$

On vérifie facilement la formule pour un polynôme constant et pour les polynômes x, x^2, x^3. Comme le second membre dépend linéairement de P, la formule est vraie encore pour une combinaison $a_0 + a_1x + a_2x^2 + a_3x^3$.

Si l'on approche une fonction f par un tel polynôme, on aura une formule simple qui donne approximativement la valeur de l'intégrale de f.

La méthode de Simpson

Soit f une fonction définie sur $[a,b]$ et soit $m = \frac{a+b}{2}$ le milieu de $[a,b]$. Approchons la fonction f par son polynôme d'interpolation P en a, m, b. Comme ce polynôme est de degré au plus 2, son intégrale sur $[a,b]$ est donnée par la formule exacte ci-dessus. Or P prend les mêmes valeurs que f en a, $\frac{a+b}{2}$ et b, donc on a la formule d'intégration approchée

$$\int_a^b f(t)\,dt \simeq \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

Erreur d'intégration. *Si f possède une dérivée quatrième continue, alors en posant $M = \max_{x\in[a,b]} \left|f^{(4)}(x)\right|$, on a*

$$\left|\int_a^b f(t)\,dt - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]\right| \leqslant \frac{(b-a)^5}{2880}M$$

Démonstration. Posons $E = \int_a^b \left[f(t)-P(t)\right] dt$, de sorte que dans la formule ci-dessus, le premier membre est $|E|$. Posons $m = \frac{a+b}{2}$ et considérons le polynôme

$$S(x) = P(x) + \frac{4}{(b-a)^2}\left[P'(m)-f'(m)\right]Q(x), \text{ où } Q(x) = (x-a)(x-m)(x-b).$$

En a,m,b, le polynôme S prend les mêmes valeurs que f. De plus, on a $Q'(m)=(m-a)(m-b)= \frac{b-a}{2}\frac{a-b}{2} = -\frac{(b-a)^2}{4}$, donc $S'(m)=f'(m)$. Le polynôme Q est de degré 3 et s'annule en a,m,b, donc $\int_a^b Q(t)\,dt = 0$ d'après la formule d'intégration exacte. Il s'ensuit que P et S ont la même intégrale sur $[a,b]$, autrement dit $E = \int_a^b \left[f(t)-S(t)\right] dt$. Posons $q(x) = (x-a)(x-m)^2(x-b)$ et écrivons l'égalité précédente sous la forme

(1) $$E = \int_a^b q(t)\,\frac{f(t)-S(t)}{q(t)}\,dt$$

Dans l'intégrale, le quotient n'est *a priori* pas défini en a, b et m. Mais la fonction $f-S$ est dérivable et vaut 0 en a, donc $\frac{f(t)-S(t)}{t-a}$ tend vers la limite $f'(a)-S'(a)$ quand t tend vers a ; en donnant cette limite comme valeur en a au quotient, on obtient une fonction continue

en a. De même, la fonction $\dfrac{f(t)-S(t)}{t-a}$ se prolonge par continuité en b. Au point m, le développement limité à l'ordre 2 de la fonction $u = f - S$ s'écrit :

$$u(t) = u(m) + u'(m)(t-m) + \frac{u''(m)}{2}(t-m)^2 + o\big[(t-m)^2\big] = \frac{u''(m)}{2}(t-m)^2 + o\big[(t-m)^2\big]$$

car $u(m) = u'(m) = 0$. On voit que $\dfrac{u(t)}{(t-m)^2}$ tend vers $(1/2)u''(m)$ quand t tend vers m, ce qui permet encore de prolonger par continuité la fonction $\dfrac{u(t)}{(t-m)^2}$ en m. Finalement, la fonction $\dfrac{f(t)-S(t)}{q(t)}$ est continue en tout point de $[a, b]$.

Les valeurs $q(t)$ étant négatives ou nulles, nous pouvons appliquer à l'intégrale (1) une proposition page 289 en choisissant $w(t) = -q(t)$. On obtient

(2) $$E = \frac{f(c) - S(c)}{q(c)} \int_a^b q(t)\,dt, \text{ où } c \text{ est un nombre dans } [a, b].$$

En raisonnant comme dans le calcul de l'erreur d'interpolation page 347, on montre que

$$f(c) - S(c) = \frac{f^{(4)}(d)}{4!} q(c), \text{ pour un certain nombre } d \in [a, b].$$

D'autre part, en intégrant le polynôme $q(t)$, on a $\displaystyle\int_a^b q(t)\,dt = -\frac{(b-a)^5}{120}$ et en reportant dans (2), il vient $E = -\dfrac{f^{(4)}(d)}{4!}\dfrac{(b-a)^5}{120}$, d'où le résultat. ■

Pratique de la méthode

Pour calculer précisément l'intégrale de f sur $[a, b]$, partageons le segment en $2n$ intervalles au moyen des points

$$x_0 = a\ ,\ x_1 = a + h\ ,\ \ldots, x_i = a + ih\ ,\ x_{2n} = a + 2nh = b\ ,\ \text{où } h = \frac{b-a}{2n}.$$

Le point x_1 est le milieu du segment $[x_0, x_2]$ et plus généralement x_{2k+1} est le milieu de $[x_{2k}, x_{2k+2}]$.

Sur l'intervalle $[x_0, x_2]$, on a la formule d'intégration approchée de Simpson

$$\int_{x_0}^{x_2} f(t)\,dt \simeq \frac{2h}{6}\big[f(x_0) + 4f(x_1) + f(x_2)\big]$$

et de même sur $[x_2, x_4]$, etc. Chaque fois, l'erreur est au plus égale à $e = \dfrac{(2h)^5}{2880}M$, où M est le maximum de $\left|f^{(4)}(x)\right|$ sur $[a, b]$. Puisqu'il y a n intervalles, l'erreur totale sur $\displaystyle\int_a^b f(t)\,dt$ est donc au plus égale à

$$ne = 2nh\frac{(2h)^4}{2880}M = (b-a)\frac{2^4}{2880}h^4 M = \frac{b-a}{180}h^4 M$$

Si l'on a pris suffisamment de points, l'erreur est en principe aussi petite qu'on veut, sous réserve des arrondis qui s'accumulent quand on fait un grand nombre d'additions.

Algorithme pour la méthode de Simpson

initialisation

a, b des nombres

p un entier positif pair

$h \leftarrow (b-a)/p$

$I_0 \leftarrow f(a) + f(b)$

$\big(I_1 \leftarrow 0\big)$, $\big(I_2 \leftarrow 0\big)$

boucle : pour $i = 1, \ldots, p{-}1$, faire

$x = a + ih$

si i est pair, alors $I_2 \leftarrow I_2 + f(x)$,

sinon $I_1 \leftarrow I_1 + f(x)$

fin : $I \leftarrow h(I_0 + 2I_2 + 4I_1)/3$

À la fin de cet algorithme, la variable I contient une valeur approchée de l'intégrale.

Exemple. La longueur de l'ellipse d'équation $x^2 + (y^2/4) = 1$ est

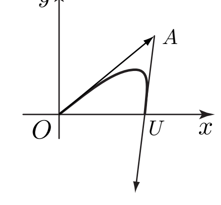
$$L = 4I = 4\int_0^{\pi/2} \sqrt{(\sin t)^2 + 4(\cos t)^2}\, dt$$

L'intégrale ne s'exprimant pas au moyen des fonctions usuelles, calculons-la de manière approchée en prenant $n = 2$. On a $a = 0$, $b = \pi/2$, $h = \pi/8$ et comme valeur approchée de I, on obtient $I \simeq 2{,}4228$.

L'erreur E est majorée par $\dfrac{b{-}a}{180}h^4M = \dfrac{\pi/2}{180}(\pi/8)^4M$, où M est de l'ordre de 40, ce qui donne $E \leqslant 8.10^{-3}$. En fait, la valeur exacte de I est proche de 2,4221, d'où une erreur vraie d'environ 7.10^{-4}. Pour la longueur de l'ellipse, on trouve comme approximation $L = 4I \simeq 9{,}69$, alors que la valeur exacte est plus proche de 9,6884.

Exercices

@ **1. Un triangle de sécurité pour les courbes splines.**



Soit $A = (a,b)$ un point du plan non situé sur l'axe des abscisses. Considérons la courbe spline passant par l'origine O en $t = 0$ avec vecteur tangent $\overrightarrow{OA}$ et par le point $U = (1,0)$ en $t = 1$ avec vecteur tangent $-\overrightarrow{UA}$. Son équation est (voir page 354)

$$\begin{cases} x(t) &= -t^3 + (2{-}a)t^2 + at \\ y(t) &= -bt^2 + bt \end{cases}, \text{ où } 0 \leqslant t \leqslant 1.$$

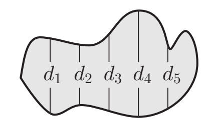
a) Montrer qu'un point $M = (x,y)$ du plan est du côté de U par rapport à la droite OA si et seulement si $bx - ay$ a le signe de b. Montrer que M est du côté de O par rapport à la droite UA si et seulement si $(a{-}1)y - bx + b$ est du signe de b.

b) Montrer que la courbe est dans le triangle OAU.

Cette propriété est utile par exemple en robotique où l'on construit des fonctions splines pour faire évoluer les paramètres d'un mouvement : des triangles de sécurité disjoints permettent d'éviter les collisions.

2. La figure ci-contre est la carte d'un étang : on mesure tous les 10 mètres les distances $d_1, d_2, \ldots, d_5$:

d_1	d_2	d_3	d_4	d_5
26 m	21,7 m	28,3 m	36,7 m	20,7 m



Montrer que ces données permettent d'estimer à 13,88 ares l'aire de l'étang.

@ **3. Orthogonalité des polynômes de Chebychev.** Les polynômes de Chebychev T_n sont définis, pour n entier positif ou nul, par la formule $T_n(\cos\theta) = \cos(n\theta)$ (page 348). Rappelons que T_n est de degré n.

a) On pose $I_{n,p} = \int_{-1}^{1} T_n(t)T_p(t)(1-t^2)^{-1/2}\,dt$. En faisant le changement de variable $t = \cos\theta$, montrer que l'on a $I_{0,0} = \pi$, $I_{n,p} = 0$ si $n \neq p$ et $I_{n,n} = \pi/2$ si $n \geqslant 1$.

b) Notons E l'espace vectoriel des fonctions continues sur $[-1, 1]$. Pour toutes fonctions f et g appartenant à E, on définit leur produit scalaire

$$f \cdot g = \int_{-1}^{1} f(t)g(t)(1-t^2)^{-1/2}\,dt.$$

Posons $U_0 = (1/\sqrt{\pi})T_0$ et $U_n = (\sqrt{2/\pi})T_n$ pour $n \geqslant 1$.

(i) Vérifier que $f \cdot g$ est un produit scalaire sur E et que les polynômes U_n forment une famille orthonormée (voir chapitre **7**).

(ii) Soit F_n le sous-espace vectoriel de E formé des polynômes de degré inférieur ou égal à n. Montrer que $U_0, U_1, \ldots, U_n$ est une base orthonormée de F_n.

c) Étant donné une fonction f continue sur $[-1,1]$, posons $c_n = f \cdot U_n$: le polynôme $P = c_0U_0 + c_1U_1 + c_2U_2 + c_3U_3$ est de degré inférieur ou égal à 3 et en notant $\|g\| = \sqrt{g \cdot g}$ la norme dans l'espace euclidien E, on a $\|f - P\| \leqslant \|f - Q\|$ pour tout autre polynôme Q de degré inférieur ou égal à 3 (page 209).
On prend $f(x) = x(\sin 2x)^2$. Montrer qu'on a à peu près $P = 0{,}756x + 1{,}66x^3$. Représenter sur un même dessin le graphe de f et celui de P entre -1 et 1.

@ **4.** On veut interpoler la fonction $f(x) = x(\sin 2x)^2$ de l'exercice précédent sur les abscisses de Chebychev $x_k = \cos\dfrac{(2k-1)\pi}{8}$, avec $k = 1, 2, 3, 4$. Montrer que le polynôme d'interpolation est à peu près $R = 0{,}387x + 0{,}629x^3$. Représenter sur un même dessin le graphe de f et celui de R entre -1 et 1.

Chapitre 12

Fonctions de plusieurs variables

1. Présentation

Il est habituel qu'une quantité dépende de plusieurs variables. Par exemple :

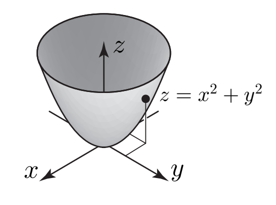
- La solution $X = \begin{bmatrix} x \\ y \end{bmatrix}$ de l'équation linéaire $\begin{bmatrix} u & -v \\ v & u \end{bmatrix} X = \begin{bmatrix} p \\ q \end{bmatrix}$ est fonction des nombres u, v, p, q.
- La température en tout point d'une pièce d'habitation est, à un instant donné, fonction des trois coordonnées spatiales du point.
- En Économie, à chaque panier constitué de n biens en quantités $x_1, x_2, \ldots, x_n$, on associe son utilité $U(x_1, x_2, \ldots, x_n)$, un nombre qui traduit la préférence du consommateur pour cette composition d'achats.

Représentation des fonctions de deux variables

Soit $u = f(x,y)$ une quantité numérique fonction des deux variables x et y. Il y a deux manières de visualiser la fonction f : par ses lignes de niveau ou par son graphe.

Les lignes de niveau. Les points (x,y) où $f(x,y)$ prend une valeur donnée k forment une courbe, appelée *ligne de niveau* de f (page 21). La ligne de niveau k a pour équation $f(x,y)=k$. Des lignes de niveaux k et k' différents n'ont aucun point commun.

Le graphe. Au dessus de chaque point $m = (x, y)$ du domaine de définition de f, plaçons dans l'espace le point M de coordonnées $\big(x, y, f(x,y)\big)$. L'ensemble de ces points M est par définition le *graphe de* f. Quand (x,y) décrit le domaine de définition D de f, le point M décrit une surface étalée sur D.



L'équation du graphe de f est $z = f(x, y)$.

Si l'on coupe la surface d'équation $z = f(x, y)$ par le plan horizontal d'équation $z = k$, on obtient la courbe formée des points (x, y, k) tels que $f(x, y) = k$: la projection de cette courbe dans le plan des x, y est la ligne de niveau k.

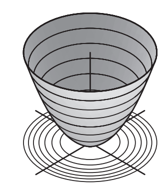


figure 1

Les lignes de niveau sont les coupes horizontales du graphe.

On peut aussi reconstituer le graphe de f en plaçant chaque ligne de niveau à sa bonne hauteur : la ligne de niveau k dans le plan d'équation $z = k$ (voir les figures du col, page 21).

► Sur la figure 1, chaque ligne de niveau $k > 0$ est un cercle, ce qui est caractéristique d'une surface de révolution d'axe Oz.

► La figure 2 montre la surface d'équation $z = \cos\frac{5\pi x}{2}\cos 3\pi y$. C'est, à un instant donné, la forme possible de la surface de l'eau dans une piscine rectangulaire où on laisse le liquide osciller librement. Sur la figure 3, on voit des lignes de niveau de cette surface.

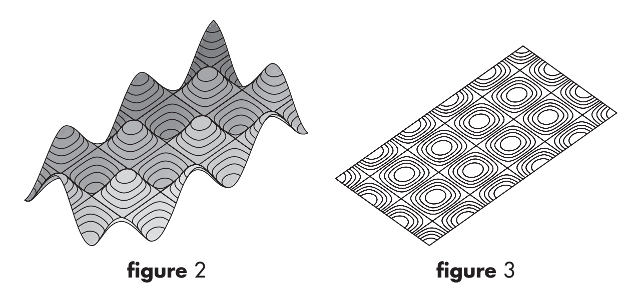


figure 2 **figure** 3

2. Normes et distances dans $\mathbb{R}^n$

On peut définir plusieurs notions de distance dans $\mathbb{R}^2$, par exemple au moyen d'un produit scalaire et de la norme associée (page 202).

► En utilisant la norme euclidienne $\|(x, y)\| = \sqrt{x^2 + y^2}$, la distance de deux points $P = (x, y)$, $P' = (x', y')$ est $d(P, P') = \|\,\overrightarrow{PP'}\,\| = \sqrt{(x'-x)^2 + (y'-y)^2}$.
Notons O le point $(0, 0)$. Si r est un nombre positif, l'ensemble des points M tels que $d(O, M) < r$ est le disque de centre O et de rayon r (bord non compris).

► On peut aussi considérer le maximum des valeurs absolues des coordonnées, en posant $\|(x, y)\|_m = \max\big(|x|, |y|\big)$. Cette fonction a bien les propriétés d'une norme (page 204), mais contrairement à la norme euclidienne, elle n'est pas associée à un produit scalaire. Pour cette norme, on définit la distance des points $P = (x, y)$ et $P' = (x', y')$ en posant

$$d_m(P, P') = \|\,\overrightarrow{PP'}\,\|_m = \max\big(|x'-x|, |y'-y|\big)$$

L'ensemble des points M tels que $d_m(O, M) < r$ est le carré de centre O dont les côtés sont parallèles aux axes et de longueur $2r$.

Les nombres $|x|$ et $|y|$ sont inférieurs ou égaux à $\sqrt{x^2 + y^2}$, donc $\|(x,y)\|_m \leqslant \|(x,y)\|$. D'autre part, $|x|$ et $|y|$ sont inférieurs ou égaux à $\|(x,y)\|_m$, donc $x^2 + y^2 \leqslant 2\|(x,y)\|_m^2$ et par suite $\|(x, y)\| \leqslant \sqrt{2}\,\|(x, y)\|_m$.

Dans $\mathbb{R}^n$, on définit de même pour $X = (x_1, x_2, \ldots, x_n)$,

la norme et la distance euclidienne :

$$\|X\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \quad , \quad d(X,Y) = \|(X-Y)\|$$

la norme et distance « du sup » :

$$\|X\|_m = \max\big(|x_1|, |x_2|, \ldots, |x_n|\big) \quad , \quad d_m(X,Y) = \|(X-Y)\|_m$$

On a les inégalités $\|X\|_m \leqslant \|X\| \leqslant \sqrt{n}\|X\|_m$. Il s'ensuit que pour des points X et Y quelconques de $\mathbb{R}^n$,

- si l'on a $d(X,Y) < a$, alors $d_m(X,Y) < a$
- si l'on a $d_m(X,Y) < b$, alors $d(X,Y) < b\sqrt{n}$.

Définitions

Soit f une fonction de n variables définie sur un domaine D et soit $A \in D$.

- On dit que $f(X)$ *tend vers* ℓ *quand* X *tend vers* A si $\big|f(X) - \ell\big|$ tend vers 0 quand $d(A,X)$ tend vers 0 ; cette propriété se note $\lim\limits_{X \to A} f(X) = \ell$.
- La fonction f est *continue en* A si $f(X)$ tend vers $f(A)$ quand X tend vers A.

Définition

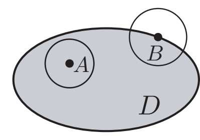
Soit D une partie de $\mathbb{R}^n$ et soit $A \in D$. On dit que A est un *point intérieur à* D s'il existe un nombre $r > 0$ tel que D contient tous les points M dont la distance à A est moindre que r.

D'après les inégalités précédentes, on peut utiliser la distance qu'on veut pour vérifier si un point est intérieur.

Propriété d'un point intérieur. *Si* $A = (a_1, a_2, \ldots, a_n)$ *est un point intérieur à* D, *alors pour tous nombres réels* $t_1, t_2, \ldots, t_n$ *assez petits, le point* $(a_1 + t_1, a_2 + t_2, \ldots, a_n + t_n)$ *appartient à* D.

Posons en effet $T = (t_1, t_2, \ldots, t_n)$, de sorte que $d_m(A, A+T) = \|(A+T) - A\|_m = \|T\|_m = \max|t_i|$. Supposons A intérieur à D, donc il existe un nombre $r > 0$ tel que $\big[d_m(A, A+T) < r \Rightarrow A+T \in D\big]$. Si tous les $|t_i|$ sont inférieurs à r, alors leur maximum est inférieur à r, donc $A+T$ appartient à D.

Intuitivement, un point A est intérieur à D quand D contient un voisinage de A. Les points du bord ne sont pas intérieurs. Par exemple, pour le domaine elliptique ci-contre, le point A est intérieur et le point B ne l'est pas.



Pour une fonction f d'une variable réelle, on sait que si f est continue sur un segment, elle y atteint un maximum et un minimum. Ce résultat est encore vrai pour une fonction de plusieurs variables continue sur un domaine qui, comme les segments, contient son bord.

On dit qu'un domaine D de $\mathbb{R}^n$ est *compact* s'il a les propriétés suivantes :

i) D est borné, autrement dit il existe un nombre R tel que l'on ait $\|X\| \leqslant R$ pour tout $X \in D$.

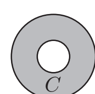
ii) D est défini par une ou plusieurs inégalités $\varphi(x_1, \ldots, x_n) \leqslant 0$, où les fonctions φ sont continues sur $\mathbb{R}^n$.

Si un point M est sur le bord de D, alors en M l'une au moins des inégalités de définition devient une égalité. Un domaine borné et contenant son bord est compact.

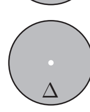
Théorème. *Soit D un domaine compact de $\mathbb{R}^n$. Si $f : D \to \mathbb{R}$ est une fonction continue, le maximum de f sur D et le minimum de f sur D existent et sont atteints en des points de D.*

Exemples

- Soit A un point dans l'espace $\mathbb{R}^n$ et soit $r > 0$. Les points M dont la distance à A est au plus r forment un domaine $D = \{M \in \mathbb{R}^n \mid d(A, M) \leqslant r\}$ compact ; si $n = 2$, D est un disque ; si $n \geqslant 3$, on dit que D est une *boule*. Le bord de D est la *sphère* formée des points M tels que $d(A, M) = r$: la sphère aussi est un domaine compact (elle est définie par les inégalités $0 \leqslant d(A, M) - r \leqslant 0$).
- Dans $\mathbb{R}^2$, la couronne $C = \{M \in \mathbb{R}^2 \mid 1 \leqslant d(A, M) \leqslant 2\}$ est compacte. Le bord de C est la réunion des deux cercles de rayon 1 et 2 centrés en A.



- Enlevons son centre A au disque $D = \{M \in \mathbb{R}^2 \mid d(A, M) \leqslant 1\}$: on obtient le domaine $\Delta = D \setminus \{A\}$ dont le bord est formé du cercle unité et du point A. Puisque A n'est pas dans Δ, le domaine Δ n'est pas compact.



3. Dérivées partielles

Soit $(x, y) \mapsto f(x, y)$ une fonction de deux variables définie sur un domaine $D \subset \mathbb{R}^2$ et soit (a, b) un point intérieur à D.

Pour tout nombre t assez petit en valeur absolue, le point $(a + t, b)$ est par hypothèse dans D. La fonction d'une variable $x \mapsto f(x, b)$ est donc définie sur un certain intervalle ouvert contenant a. Si cette fonction est dérivable en a, sa dérivée en a est $\lim_{t \to 0} \dfrac{f(a + t, b) - f(a, b)}{t}$.

Définition

Si la fonction $x \mapsto f(x, b)$ est dérivable en a, sa dérivée se note $\dfrac{\partial f}{\partial x}(a, b)$ et s'appelle la *dérivée partielle de f par rapport à x* au point (a, b). De même, la dérivée en b de la fonction $y \mapsto f(a, y)$ s'appelle la *dérivée partielle de f par rapport à y* au point (a, b) et se note $\dfrac{\partial f}{\partial y}(a, b)$.

Par définition,

$$\frac{\partial f}{\partial x}(a,b) = \lim_{t\to 0}\frac{f(a+t,b)-f(a,b)}{t} \quad \text{et} \quad \frac{\partial f}{\partial y}(a,b) = \lim_{t\to 0}\frac{f(a,b+t)-f(a,b)}{t}.$$

Pour une fonction f de n variables $(x_1, x_2, \ldots, x_n)$, on définit de même la dérivée partielle $\dfrac{\partial f}{\partial x_i}$ par rapport à la i-ème variable.

Pour calculer la dérivée partielle par rapport à x_i en un point $A = (a_1, a_2, \ldots, a_n)$, on fixe toutes les variables sauf x_i en leur donnant leur valeur en A et l'on a

$$\frac{\partial f}{\partial x_i}(A) = \lim_{t\to 0}\frac{f(a_1,\ldots,a_i+t,\ldots,a_n)-f(a_1,\ldots,a_n)}{t}.$$

3.1 Approximation affine, différentielle

Comme pour une fonction d'une variable, cherchons à estimer la différence $f(X)-f(A)$ quand X est voisin de A et qu'on néglige les quantités infiniment petites devant la distance $d(A,X) = \|X-A\|$.

Proposition. *Soit f une fonction de deux variables ayant des dérivées partielles continues en un point $A=(a,b)$. Alors pour tout $X=(x,y)$ assez voisin de A, on a*

$$f(X) = f(A) + (x-a)\frac{\partial f}{\partial x}(A) + (y-b)\frac{\partial f}{\partial y}(A) + o\big(\|X-A\|\big)$$

Démonstration. Considérons un petit rectangle A, P, X, Q à côtés parallèles aux axes et contenu dans D et posons $X=(a+h,b+k)$, $P=(a+h,b)$ et $Q=(a,b+k)$. D'après le théorème des accroissements finis pour les fonctions d'une variable, on a

$$f(P)-f(A) = f(a+h,b)-f(a,b) = h\frac{\partial f}{\partial x}(U)\ ,\ \text{où } U \text{ est sur le segment } AP$$

$$f(X)-f(P) = f(a+h,b+k)-f(a+h,b) = k\frac{\partial f}{\partial y}(V)\ ,\ \text{où } V \text{ est sur le segment } PX$$

En ajoutant ces deux égalités, on obtient :

(1) $$f(X)-f(A) = h\frac{\partial f}{\partial x}(U) + k\frac{\partial f}{\partial y}(V)$$

Quand h et k tendent vers 0, les points U et V tendent vers A. Si les fonctions dérivées partielles $\dfrac{\partial f}{\partial x}$ et $\dfrac{\partial f}{\partial y}$ sont continues en A, alors $\dfrac{\partial f}{\partial x}(U)$ tend vers $\dfrac{\partial f}{\partial x}(A)$ et $\dfrac{\partial f}{\partial y}(V)$ tend vers $\dfrac{\partial f}{\partial y}(A)$, ce qu'on peut écrire sous la forme

$$\frac{\partial f}{\partial x}(U) = \frac{\partial f}{\partial x}(A) + \varepsilon_1(h,k) \quad \text{et} \quad \frac{\partial f}{\partial y}(V) = \frac{\partial f}{\partial y}(A) + \varepsilon_2(h,k)$$

où $\varepsilon_1(h,k)$ et $\varepsilon_2(h,k)$ tendent vers 0 quand h et k tendent vers 0. L'égalité (1) devient

$$f(X)-f(A) = h\frac{\partial f}{\partial x}(A) + k\frac{\partial f}{\partial y}(A) + h\varepsilon_1 + k\varepsilon_2\,.$$

Montrons que la quantité $r(h,k) = h\varepsilon_1 + k\varepsilon_2$ est négligeable devant la norme de (h,k) quand h et k tendent vers 0. Puisque $|h|$ et $|k|$ sont inférieurs ou égaux à $\|(h,k)\|_m$, on a $\left|r(h,k)\right| \leqslant |\varepsilon_1||h|+|\varepsilon_2||k| \leqslant \left[|\varepsilon_1|+|\varepsilon_2|\right]\|(h,k)\|_m$. Le rapport $\dfrac{r(h,k)}{\|(h,k)\|_m}$ est inférieur à $|\varepsilon_1(h,k)|+|\varepsilon_2(h,k)|$ qui tend vers 0 quand h et k tendent vers 0 : on a donc $r(h,k) \underset{(h,k)\to(0,0)}{\ll} \|(h,k)\|_m$. ■

Si f est une fonction de n variables à dérivées partielles continues et si $A=(a_1,\ldots,a_n)$, on a de même pour tout $X = (h_1, \ldots, h_n)$ assez voisin de A

$$f(X)=f(A)+(x_1-a_1)\frac{\partial f}{\partial x_1}(A)+(x_2-a_2)\frac{\partial f}{\partial x_2}(A)+\cdots+(x_n-a_n)\frac{\partial f}{\partial x_n}(A)+o\big(\|X-A\|\big)$$

On dit que $f(A) + (x_1-a_1)\dfrac{\partial f}{\partial x_1}(A) + (x_2-a_2)\dfrac{\partial f}{\partial x_2}(A) + \cdots + (x_n-a_n)\dfrac{\partial f}{\partial x_n}(A)$ est *l'approximation affine* de f au point A.

Dans la suite, nous supposerons toujours que les fonctions considérées ont des dérivées partielles continues.

Définitions

Soit $z = f(x_1, x_2, \ldots, x_n)$ une fonction de n variables.

- La *différentielle de f en A* est la fonction linéaire
  $dz = \dfrac{\partial f}{\partial x_1}(A)\,dx_1 + \dfrac{\partial f}{\partial x_2}(A)\,dx_2 + \cdots + \dfrac{\partial f}{\partial x_n}(A)\,dx_n$ de variables $dx_1, \ldots, dx_n$.
- La matrice-ligne $J_f(A) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}(A) & \dfrac{\partial f}{\partial x_2}(A) & \cdots & \dfrac{\partial f}{\partial x_n}(A) \end{bmatrix}$ s'appelle la *matrice jacobienne* de f en A.

En posant $dX = \begin{bmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{bmatrix}$, il vient $dz = J_f(A)\,dX$:

la matrice $J_f(A)$ est la matrice de la différentielle de f en A.

Si l'on donne aux variables $dx_1,\ldots,dx_n$ des valeurs assez petites, alors $dz=J_f(A)dX$ est une approximation au premier ordre de la différence $f(X) - f(A)$.

Le vecteur gradient

Il est intéressant d'introduire le vecteur-colonne ${}^t\big(J_f(A)\big)$, car $J_f(A)dX$ est alors le produit scalaire de ce vecteur et de dX.

Définition

Le *gradient* de f en A est le vecteur $\overrightarrow{\mathrm{Grad}}_f(A) = \Big(\dfrac{\partial f}{\partial x_1}(A)\,,\ \dfrac{\partial f}{\partial x_2}(A)\,,\ \ldots\ ,\ \dfrac{\partial f}{\partial x_n}(A)\Big)$.

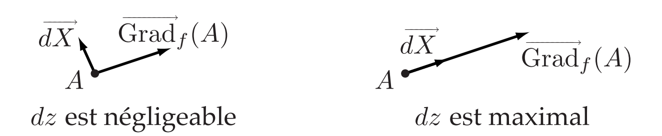
La différentielle s'écrit ainsi $dz = \overrightarrow{\mathrm{Grad}}_f(A) \cdot \overrightarrow{dX}$, où $(\ \cdot\ )$ désigne le produit scalaire euclidien.

On sait que le produit scalaire est maximum quand les deux vecteurs sont colinéaires et de même sens, et que le produit scalaire est nul s'ils sont orthogonaux.

Pour une petite variation vectorielle $\overrightarrow{dX}$, de norme donnée, appliquée au point A,

► *dz est maximum et positif quand $\overrightarrow{dX}$ est dans la direction du gradient $\overrightarrow{\text{Grad}}_f(A)$,*

► *dz est négligeable devant $\| \overrightarrow{dX} \|$ quand $\overrightarrow{dX}$ est orthogonal à $\overrightarrow{\text{Grad}}_f(A)$.*



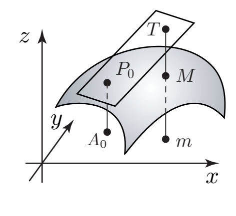
Plan tangent à une surface

Soit S la surface d'équation $z = f(x, y)$, où f est définie sur un domaine D. Soit $A_0 = (x_0, y_0)$ un point de D et soit $P_0 = \big(x_0, y_0, z_0\big)$ le point de S correspondant, donc $z_0 = f(x_0, y_0)$.

Définition

Si la matrice jacobienne $J_f(A_0) = \begin{bmatrix} \dfrac{\partial f}{\partial x}(A_0) & \dfrac{\partial f}{\partial y}(A_0) \end{bmatrix}$ n'est pas nulle, le plan d'équation $z - z_0 = J_f(A_0)\begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}$ s'appelle le *plan tangent* à S au point P_0.

L'équation du plan tangent est donc $z - z_0 = (x - x_0)\dfrac{\partial f}{\partial x}(x_0, y_0) + (y - y_0)\dfrac{\partial f}{\partial y}(x_0, y_0)$.



Considérons un point $m = (x, y)$ voisin de A_0. Sur une même verticale menée par m, se trouvent le point $M = \big(x, y, z\big)$ appartenant à S et le point $T = (x, y, t)$ appartenant au plan tangent à S en P. On a $z = f(x, y)$ et $t = z_0 + J_f(A_0)\begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}$, d'où

$$\begin{aligned} \overline{MT} &= z-t \\ &= f(x,y)-f(x_0,y_0) - (x-x_0)\frac{\partial f}{\partial x}(x_0,y_0) - (y-y_0)\frac{\partial f}{\partial y}(x_0,y_0) \end{aligned}$$

Par la proposition précédente, cette quantité est négligeable devant $A_0m = \|(x-x_0, y-y_0)\|$ quand m tend vers A_0.

► Le plan tangent est horizontal (c'est-à-dire d'équation $z = z_0$) si et seulement si $\dfrac{\partial f}{\partial x}(x_0, y_0) = \dfrac{\partial f}{\partial y}(x_0, y_0) = 0$. Cela se produit notamment si la surface présente un maximum ou un minimum en (x_0, y_0, z_0).

3.2 Calcul des différentielles

Différentielle d'une somme ou d'un produit

Si $u = f(x_1,\dots,x_n)$ et $v = g(x_1,\dots,x_n)$ sont des fonctions de n variables définies sur un même domaine D, alors pour les matrices jacobiennes en un point $A \in D$, on a

$$J_{f+g}(A) = J_f(A) + J_g(A) \quad \text{et} \quad J_{fg}(A) = f(A)J_g(A) + g(A)J_f(A)$$

Différentielle d'une composée

1) Supposons que $z = f(x_1,\dots,x_n)$ et que $x_1,\dots,x_n$ sont des fonctions $x_1(t),\dots,x_n(t)$ d'une variable réelle t. La différentielle de x_i en t_0 est $dx_i = x_i'(t_0)dt$ et en posant $M_0 = \big(x_1(t_0),\dots,x_n(t_0)\big)$, la différentielle de z en M_0 est

$$\frac{\partial f}{\partial x_1}(M_0)dx_1 + \frac{\partial f}{\partial x_2}(M_0)dx_2 + \cdots + \frac{\partial f}{\partial x_n}(M_0)dx_n\,.$$

La différentielle en t_0 de $z = f\big(x_1(t),\dots,x_n(t)\big)$ est donc

$$dz = \left(\frac{\partial f}{\partial x_1}(M_0)x_1'(t_0) + \frac{\partial f}{\partial x_2}(M_0)x_2'(t_0) + \cdots + \frac{\partial f}{\partial x_n}(M_0)x_n'(t_0)\right)dt$$

On en déduit

$$\frac{dz}{dt}(t_0) = \frac{\partial f}{\partial x_1}(M_0)x_1'(t_0) + \frac{\partial f}{\partial x_2}(M_0)x_2'(t_0) + \cdots + \frac{\partial f}{\partial x_n}(M_0)x_n'(t_0) = \overrightarrow{\mathrm{Grad}}_f(M_0)\cdot\begin{bmatrix} x_1'(t_0) \\ \vdots \\ x_n'(t_0)\end{bmatrix}$$

2) Soit $z = f(x,y)$ une fonction de deux variables. Supposons que chacune des quantités x et y est fonction de deux variables u, v : on a $x(u,v)$ et $y(u,v)$. En $U_0 = (u_0, v_0)$, x et y ont pour différentielles

$$dx = \frac{\partial x}{\partial u}(U_0)du + \frac{\partial x}{\partial v}(U_0)dv$$
$$dy = \frac{\partial y}{\partial u}(U_0)du + \frac{\partial y}{\partial v}(U_0)dv$$

En $M_0 = (x_0, y_0) = \big(x(u_0,v_0), y(u_0,v_0)\big)$, la différentielle de z est

$$dz = \frac{\partial f}{\partial x}(M_0)dx + \frac{\partial f}{\partial y}(M_0)dy$$

donc en U_0, la différentielle de z comme fonction de u, v est

$$\begin{aligned} dz &= \frac{\partial f}{\partial x}(M_0)\left[\frac{\partial x}{\partial u}(U_0)du + \frac{\partial x}{\partial v}(U_0)dv\right] + \frac{\partial f}{\partial y}(M_0)\left[\frac{\partial y}{\partial u}(U_0)du + \frac{\partial y}{\partial v}(U_0)dv\right] \\ &= \left[\frac{\partial f}{\partial x}(M_0)\frac{\partial x}{\partial u}(U_0) + \frac{\partial f}{\partial y}(M_0)\frac{\partial y}{\partial u}(U_0)\right]du + \left[\frac{\partial f}{\partial x}(M_0)\frac{\partial x}{\partial v}(U_0) + \frac{\partial f}{\partial y}(M_0)\frac{\partial y}{\partial v}(U_0)\right]dv \end{aligned}$$

On en déduit les formules pour les dérivées partielles de z par rapport à u et à v :

$$\frac{\partial z}{\partial u} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial u} \quad \text{et} \quad \frac{\partial z}{\partial v} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v}$$

Utilisation des matrices jacobiennes

➤ Reprenons le calcul précédent en introduisant la fonction $\varphi:(u,v)\mapsto\big(x(u,v),y(u,v)\big)$ qui a deux variables et prend ses valeurs dans $\mathbb{R}^2$.

- La matrice jacobienne de f en M_0 est $J_f(M_0)=\left[\ \dfrac{\partial f}{\partial x}(M_0)\ \ \dfrac{\partial f}{\partial y}(M_0)\ \right]$.
- La fonction $(u,v)\mapsto f\big(x(u,v),y(u,v)\big)$ est par définition la composée $f\circ\varphi$ et sa matrice jacobienne en U_0 est $J_{f\circ\varphi}(U_0)=\left[\ \dfrac{\partial z}{\partial u}(U_0)\ \ \dfrac{\partial z}{\partial v}(U_0)\ \right]$, où $z=f\circ\varphi(u,v)$.

En écrivant les formules ci-dessus sous la forme matricielle

$$\left[\ \frac{\partial z}{\partial u}\ \ \frac{\partial z}{\partial v}\ \right]=\left[\ \frac{\partial f}{\partial x}\ \ \frac{\partial f}{\partial y}\ \right]\begin{bmatrix}\dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v}\\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v}\end{bmatrix}$$

on obtient

$$J_{f\circ\varphi}(U_0)=J_f(M_0)\begin{bmatrix}\dfrac{\partial x}{\partial u}(U_0) & \dfrac{\partial x}{\partial v}(U_0)\\ \dfrac{\partial y}{\partial u}(U_0) & \dfrac{\partial y}{\partial v}(U_0)\end{bmatrix}$$

La matrice $\begin{bmatrix}\dfrac{\partial x}{\partial u}(U_0) & \dfrac{\partial x}{\partial v}(U_0)\\ \dfrac{\partial y}{\partial u}(U_0) & \dfrac{\partial y}{\partial v}(U_0)\end{bmatrix}$ s'appelle la *matrice jacobienne* de φ au point U_0 ; on la note $J_\varphi(U_0)$ et notre relation s'écrit simplement

$$J_{f\circ\varphi}(U_0)=J_f(M_0)J_\varphi(U_0)\ ,\ \text{ où } M_0=\varphi(U_0)$$

Dans la matrice jacobienne de φ, la première ligne est la matrice jacobienne J_x de la fonction $(u,v)\mapsto x(u,v)$ et la deuxième ligne est J_y, matrice jacobienne de $(u,v)\mapsto y(u,v)$.

$$\text{format de la matrice } J_\varphi\ :\quad \begin{array}{c} \\ x\\ y\end{array}\begin{array}{cc}\frac{\partial}{\partial u} & \frac{\partial}{\partial v}\\ \left[\begin{array}{cc}\cdots & \cdots\\ \cdots & \cdots\end{array}\right. & \left.\begin{array}{c}\\ \\ \end{array}\right]\end{array}$$

➤ Prenons le cas d'une fonction $f:(u,v)\mapsto\big(x(u,v),y(u,v),z(u,v)\big)$ à deux variables et prenant ses valeurs dans $\mathbb{R}^3$. La matrice jacobienne de f est de même

$$J_f=\begin{bmatrix}J_x\\ J_y\\ J_z\end{bmatrix},\ \text{ où } J_x, \text{ par exemple, est la ligne } \left[\ \frac{\partial x}{\partial u}\ \ \frac{\partial x}{\partial v}\ \right].$$

La matrice jacobienne de f a trois lignes et deux colonnes.

Ces calculs se généralisent à un nombre quelconque de variables.

Règle. *Pour calculer les dérivées partielles d'une fonction composée $f\circ g$, on calcule sa matrice jacobienne $J_{f\circ g}(U)=J_f\big(g(U)\big)J_g(U)$.*

Changement de coordonnées

Coordonnées polaires. Donnons-nous un repère orthonormé $(O;\vec{i},\vec{j})$ du plan. Tout point $M=(x,y)$ différent de l'origine a des coordonnés polaires (page 39)

$$r = d(O,M) = \sqrt{x^2+y^2} \quad \text{et} \quad \theta = \widehat{\vec{i},\overrightarrow{OM}} \;, \quad \text{où} \quad 0 \leqslant \theta < 2\pi.$$

On calcule x et y au moyen de r et θ par les formules simples $x=r\cos\theta$, $y=r\sin\theta$. Si une quantité w est fonction du point M, on peut l'exprimer comme une fonction des coordonnées cartésiennes ou comme une fonction des coordonnées polaires. On a

$$dx = \cos\theta\, dr - r\sin\theta\, d\theta = \frac{x}{r} dr - y\, d\theta$$
$$dy = \sin\theta\, dr + r\cos\theta\, d\theta = \frac{y}{r} dr + x\, d\theta$$

donc pour une fonction $w = f(x,y) = g(r,\theta)$, il vient

$$\begin{bmatrix} \dfrac{\partial w}{\partial r} & \dfrac{\partial w}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial w}{\partial x} & \dfrac{\partial w}{\partial y} \end{bmatrix} \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = \begin{bmatrix} \dfrac{\partial w}{\partial x} & \dfrac{\partial w}{\partial y} \end{bmatrix} \begin{bmatrix} x/r & -y \\ y/r & x \end{bmatrix}$$

En inversant la matrice carrée, on obtient

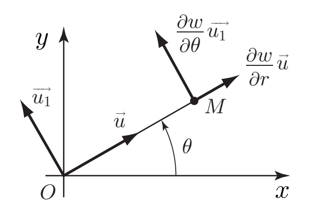
$$\begin{bmatrix} \dfrac{\partial w}{\partial x} & \dfrac{\partial w}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial w}{\partial r} & \dfrac{\partial w}{\partial \theta} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -(\sin\theta)/r & (\cos\theta)/r \end{bmatrix} = \begin{bmatrix} \dfrac{\partial w}{\partial r} & \dfrac{\partial w}{\partial \theta} \end{bmatrix} \begin{bmatrix} x/r & y/r \\ -y/r^2 & x/r^2 \end{bmatrix}$$

Expression du gradient en coordonnées polaires. Soit $\vec{u} = \dfrac{1}{\| \overrightarrow{OM} \|}\overrightarrow{OM}$ le vecteur unitaire dans la direction de $\overrightarrow{OM}$ et soit $\vec{u_1}$ le vecteur unitaire directement orthogonal à $\vec{u}$. On a

$$\vec{u} = \cos\theta\,\vec{i} + \sin\theta\vec{j}$$
$$\vec{u_1} = \cos\left(\theta + \frac{\pi}{2}\right)\vec{i} + \sin\left(\theta + \frac{\pi}{2}\right)\vec{j} = -\sin\theta\,\vec{i} + \cos\theta\vec{j}$$

et dans les coordonnées polaires, le gradient s'exprime par :

$$\overrightarrow{\text{Grad}}_w = \frac{\partial w}{\partial r}\,\vec{u} + \frac{1}{r}\frac{\partial w}{\partial \theta}\vec{u_1}$$



Coordonnées sphériques. Soit M un point de l'espace. Projetons M en m' sur le plan xOy et en m'' sur l'axe Oz. Soit $r = d(O,M)$ et si M n'est pas sur l'axe Oz, posons

$$\theta = \widehat{\vec{i},\overrightarrow{Om'}} \quad (0 \leqslant \theta < 2\pi) \quad \text{et} \quad \varphi = \widehat{\overrightarrow{Om'},\overrightarrow{OM}} \quad (-\pi/2 < \varphi < \pi/2)$$

Puisque $Om' = r\cos\varphi$, le point m' a pour coordonnées polaires $(r\cos\phi, \theta)$ et l'on a $\overline{Om''} = OM\sin\varphi$. Les nombres

$$r = \sqrt{x^2+y^2+z^2}, \quad \theta, \quad \varphi$$

s'appellent les *coordonnées sphériques* de M.

Les coordonnées cartésiennes (x,y,z) de M sont données par les formules

$$x = r\cos\theta\cos\varphi \quad , \quad y = r\sin\theta\cos\varphi \quad \text{et} \quad z = r\sin\varphi$$

Pour une fonction $w = f(x, y, z) = g(r, \theta, \varphi)$, il vient

$$\begin{bmatrix} \frac{\partial w}{\partial r} & \frac{\partial w}{\partial \theta} & \frac{\partial w}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} & \frac{\partial w}{\partial z} \end{bmatrix} \begin{bmatrix} \cos\theta\cos\varphi & -r\sin\theta\cos\varphi & -r\cos\theta\sin\varphi \\ \sin\theta\cos\varphi & r\cos\theta\cos\varphi & -r\sin\theta\sin\varphi \\ \sin\varphi & 0 & r\cos\varphi \end{bmatrix}$$

Pour écrire la matrice de droite, on a disposé en ligne les matrices jacobiennes de x, de y et de z, chacune de ces coordonnées étant fonction des variables r, θ, φ.

3.3 Applications aux variations d'une fonction

Une fonction constante a toutes ses dérivées partielles identiquement nulles et la réciproque est vraie si le domaine D est, comme un intervalle, d'un seul morceau. Soit f une fonction définie sur un domaine $D \subset \mathbb{R}^n$.

Caractérisation des fonctions constantes. *Supposons que dans le domaine D, deux points quelconques peuvent toujours être joints par une courbe paramétrée dérivable. Alors f est constante si et seulement si toutes ses dérivées partielles sont nulles sur D.*

Démonstration. Supposons par exemple que la fonction a deux variables et soit $A=(a,b)$ un point intérieur à D. Si f est constante, alors $g(t)=f(t,b)$ est constante, donc $g'(a)=\frac{\partial f}{\partial x}(A)=0$; de même, on a $\frac{\partial f}{\partial y}(A) = 0$. Réciproquement, supposons que les dérivées partielles de f sont identiquement nulles. Pour tout point $A_1 = (a_1, b_1)$ appartenant à D, il existe par hypothèse une courbe paramétrée $M(t) = \big(x(t), y(t)\big)$ telle que $\big(x(t_0), y(t_0)\big) = A$ et $\big(x(t_1), y(t_1)\big) = A_1$. Posons $h(t) = f\big(x(t), y(t)\big)$, de sorte que $h(t_0) = f(A)$ et $h(t_1) = f(A_1)$. On a $h'(t) = \frac{\partial f}{\partial x}\big(M(t)\big)x'(t) + \frac{\partial f}{\partial y}\big(M(t)\big)y'(t)$ Puisque les dérivées partielles sont nulles, $h'(t) = 0$ quel que soit t, donc h est constante et en particulier $f(A) = h(t_0) = h(t_1) = f(A_1)$. ■

Fonction constante par rapport à une variable

Soit f une fonction de deux variables sur un domaine D. Faisons les hypothèses suivantes :

a) la dérivée partielle $\frac{\partial f}{\partial x}$ est identiquement nulle,

b) si deux points de D sont situés sur une même parallèle à Ox, le segment qui les joint est dans le domaine.

Alors les valeurs $f(x, y)$ ne dépendent pas de x, autrement dit $f(x, y) = g(y)$, où g est une fonction de la seule variable y.

L'hypothèse (b) est vérifiée si par exemple D est le plan tout entier, ou une bande horizontale, ou un demi-plan de frontière parallèle à Ox.

Soient en effet (a, b) et (a', b) des point de D situés sur une même parallèle à Ox. Si t est entre a et a', le point (t, b) est par hypothèse dans D. La fonction $g(t) = f(t, b)$ est donc définie sur le segment d'extrémités a et a' et puisque la première dérivée partielle de f est nulle, on a $g'(t) = \frac{\partial f}{\partial x}(t, b) = 0$. Il s'ensuit que g est constante entre a et a',

autrement dit $f(a,b) = g(a) = g(a') = f(a',b)$. Ainsi f prend la même valeur en tous les points situés sur la parallèle à Ox menée par (a,b) et cette valeur ne dépend que de b.

Exemple : fonction sphérique. Soit $w = f(x,y,z)$ une fonction du point $M = (x,y,z)$ de l'espace. Cherchons à quelle condition w ne dépend que de la distance $r = d(O,M)$.
On doit avoir $w = g(r)$. Puisque $r^2 = x^2+y^2+z^2$, il vient $r\,dr = x\,dx+y\,dy+z\,dz$, d'où

$$\frac{\partial w}{\partial x} = \frac{dw}{dr}\frac{\partial r}{\partial x} = \frac{x}{r}g'(r) \quad , \quad \frac{\partial w}{\partial y} = \frac{dw}{dr}\frac{\partial r}{\partial y} = \frac{y}{r}g'(r) \quad , \quad \frac{\partial w}{\partial z} = \frac{dw}{dr}\frac{\partial r}{\partial z} = \frac{z}{r}g'(r)$$

autrement dit $\overrightarrow{\mathrm{Grad}}_w(x,y,z) = \dfrac{g'(r)}{r}\,[\,x\;\;y\;\;z\,]$. Ainsi le gradient de w en M doit être colinéaire à $\overrightarrow{OM}$.

Réciproquement, supposons qu'en tout point M, le gradient $\left(\frac{\partial w}{\partial x}, \frac{\partial w}{\partial y}, \frac{\partial w}{\partial z}\right)$ est colinéaire à $\overrightarrow{OM}$ et que si deux points sont sur une même droite issue de l'origine, le segment qui les joint est dans le domaine de définition de f (sur une telle droite, θ et φ sont constants). Rappelons que $\dfrac{\partial x}{\partial \theta} = -r\sin\theta\cos\varphi = -y$, $\dfrac{\partial y}{\partial \theta} = r\cos\theta\cos\varphi = x$ et $\dfrac{\partial z}{\partial \theta} = 0$. En posant $\left[\,\frac{\partial w}{\partial x}\;\;\frac{\partial w}{\partial y}\;\;\frac{\partial w}{\partial z}\,\right] = \lambda\,[\,x\;\;y\;\;z\,]$, on a alors

$$\frac{\partial w}{\partial \theta} = \frac{\partial w}{\partial x}(-y) + \frac{\partial w}{\partial y}x + \frac{\partial w}{\partial z}0 = (\lambda x)(-y) + (\lambda y)x = 0\,.$$

On vérifie de même que l'on a $\dfrac{\partial w}{\partial \varphi} = 0$. Par conséquent, w ne dépend que de la variable r.

Dans un bon domaine de définition, w ne dépend que de r si et seulement si, en tout point M, le gradient de w est colinéaire à $\overrightarrow{OM}$.

Vecteur Gradient et ligne de niveau

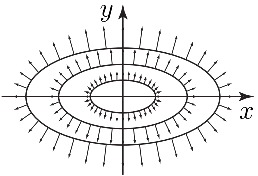
Soit $f(x,y)$ une fonction de deux variables et soit C la ligne de niveau passant par un point $M_0 = (x_0,y_0)$ donné. L'équation de C est $f(x,y) = k$, où $k = f(x_0,y_0)$. Supposons qu'au voisinage de M_0, on peut paramétrer la courbe C par $M(t) = \big(x(t), y(t)\big)$. Le point M_0 correspond à une valeur t_0 du paramètre, de sorte que $\big(x(t_0),y(t_0)\big) = (x_0,y_0)$ et $f\big(x(t),y(t)\big) = k$ pour tout t voisin de t_0.

La fonction $t \mapsto f\big(x(t),y(t)\big)$ a pour dérivée $\overrightarrow{\mathrm{Grad}}_f(M_0)\cdot\begin{bmatrix} x'(t_0) \\ y'(t_0)\end{bmatrix} = \overrightarrow{\mathrm{Grad}}_f(M_0)\cdot\dfrac{d\overrightarrow{OM}}{dt}(t_0)$, où le vecteur $\dfrac{d\overrightarrow{OM}}{dt}(t_0)$ est tangent à C en M_0. Puisque cette fonction est constante au voisinage de t_0, sa dérivée en t_0 est nulle, donc $\overrightarrow{\mathrm{Grad}}_f(M_0)\cdot\dfrac{d\overrightarrow{OM}}{dt}(t_0) = 0$.

► *En tout point, le vecteur gradient est orthogonal à la ligne de niveau.*

► *Si $\overrightarrow{\mathrm{Grad}}_f(M_0) \neq 0$, l'équation de la tangente en M_0 à la courbe d'équation $f(x,y) = k$ est*

$$\frac{\partial f}{\partial x}(M_0)(x-x_0) + \frac{\partial f}{\partial y}(M_0)(y-y_0) = 0$$



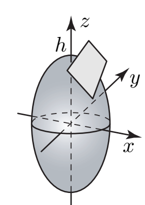
Rappelons que le vecteur gradient en M pointe vers les niveaux supérieurs à celui de M (page 365).
Sur la figure ci-contre, on a représenté, pour différentes valeurs de k, les ellipses d'équation $(1/2)x^2 + 2y^2 = k^2$. Ce sont des lignes de niveau de la fonction $f(x,y) = (1/2)x^2 + 2y^2$ et l'on voit des vecteurs $\overrightarrow{\text{Grad}}_f(x,y) = (x,4y)$ en quelques points. En un point (x_0,y_0) de l'une des ellipses, la tangente est orthogonale au gradient, donc a pour pente $\dfrac{-x_0}{4y_0}$, si $y_0 \neq 0$.

Surface de niveau. Soit $f(x,y,z)$ une fonction de trois variables et soit $M_0 = (x_0, y_0, z_0)$. L'ensemble des points $M = (x,y,z)$ tels que $f(M) = f(M_0)$ est en général une surface S passant par M_0 : c'est la surface de niveau $k = f(M_0)$ et son équation est $f(x,y,z) = k$.
Soit $\big(x(t), y(t), z(t)\big)$ une courbe paramétrée dérivable tracée sur S et passant au point M_0 pour la valeur t_0 du paramètre. On a $f\big(x(t), y(t), z(t)\big) = k$ pour tout t et en dérivant, il vient comme ci-dessus $\overrightarrow{\text{Grad}}_f(M_0) \cdot \dfrac{d\overrightarrow{OM}}{dt}(t_0) = 0$. Le vecteur gradient en M_0 est donc orthogonal à tous les vecteurs tangents à la surface en M_0, c'est-à-dire au plan tangent en M_0.

► *En tout point, le vecteur gradient est orthogonal à la surface niveau.*

► *Si* $\overrightarrow{\text{Grad}}_f(M_0) \neq 0$, *l'équation du plan tangent en* M_0 *à la surface d'équation* $f(x,y,z) = k$ *est* $\dfrac{\partial f}{\partial x}(M_0)(x-x_0) + \dfrac{\partial f}{\partial y}(M_0)(y-y_0) + \dfrac{\partial f}{\partial z}(M_0)(z-z_0) = 0.$

Exemple



L'ellipsoïde de révolution d'axe Oz, de rayon r dans le plan xOy et de hauteur $2h$ a pour équation $x^2 + y^2 + a^2z^2 = r^2$, où $a = r/h$. L'équation de son plan tangent au point $M_0 = (x_0, y_0, z_0)$ est

$$x_0(x-x_0) + y_0(y-y_0) + a^2 z_0(z-z_0) = 0\,.$$

3.4 Dérivées secondes

Soit $f : (x,y) \mapsto f(x,y)$ une fonction à valeurs réelles. La dérivée partielle $\dfrac{\partial f}{\partial x}$ est encore une fonction des deux variables x et y, donc est susceptible d'avoir des dérivées partielles. Ce sont des dérivées partielles secondes, que l'on note

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} \quad \text{et} \quad \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x}$$

De même, il peut exister des dérivées partielles secondes

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y} \quad \text{et} \quad \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2}$$

En fait, l'ordre dans lequel on fait les dérivations ne compte pas, pourvu que la fonction f soit assez régulière (théorème ci-dessous) : dans ce cas, les dérivées mixtes

$\dfrac{\partial^2 f}{\partial x \partial y}$ et $\dfrac{\partial^2 f}{\partial y \partial x}$ sont égales et il n'y a que trois dérivées partielles secondes : $\dfrac{\partial^2 f}{\partial x^2}$, $\dfrac{\partial^2 f}{\partial y^2}$ et $\dfrac{\partial^2 f}{\partial y \partial x} = \dfrac{\partial^2 f}{\partial x \partial y}$.

Exemple. Par exemple, pour $f(x, y) = x^2 y + 2xy^3$, on a

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left[x^2 + 2x(3y^2)\right] = 2x + 6y^2 \quad \text{et} \quad \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}[2xy + 2y^3] = 2x + 6y^2$$

Théorème de Schwarz. *Si la fonction* $f(x_1, x_2, \ldots, x_n)$ *a ses dérivées partielles secondes continues, alors* $\dfrac{\partial^2 f}{\partial x_i \partial x_j} = \dfrac{\partial^2 f}{\partial x_j \partial x_i}$.

Laplacien et fonctions harmoniques

En Physique, la recherche d'un potentiel conduit souvent à des fonctions $f(x, y, z)$ telles que $\Delta f = 0$, où $\Delta f = \dfrac{\partial^2 f}{\partial x^2} + \dfrac{\partial^2 f}{\partial y^2} + \dfrac{\partial^2 f}{\partial z^2}$ est le *laplacien* : on les appelle des fonctions *harmoniques*.

Exemples de fonctions harmoniques à deux variables

- Les fonctions $e^{ax}\cos(\omega y)$ et $e^{ax}\sin(\omega y)$ sont harmoniques.
- Dans les coordonnées polaires (r, θ), l'expression du laplacien de f est

$$\Delta f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2}$$

  Au moyen de cette formule, on vérifie que si k est une constante, la fonction $k \ln r$ est harmonique dans le plan privé de l'origine ; de même, si a et b sont des constantes, la fonction $a\theta + b$ est harmonique dans le domaine défini par $r > 0$, $0 < \theta < 2\pi$.
- On utilise aussi souvent les fonctions harmoniques $r^k \cos(k\theta)$ et $r^k \sin(k\theta)$, où k est un entier positif ou négatif. Comme une somme de fonctions harmoniques est harmonique, les fonctions $\sum_{k=-n}^{p} r^k \cos(k\theta + \varphi_k)$ sont harmoniques.

Les figures ci-dessous montrent des lignes de niveau pour les fonctions harmoniques $V_{-1}(r, \theta) = (1/r)\cos\theta$, $V_2(r, \theta) = r^2 \cos(2\theta + \pi/2)$ et $V_3(r, \theta) = r^3 \cos(3\theta)$.

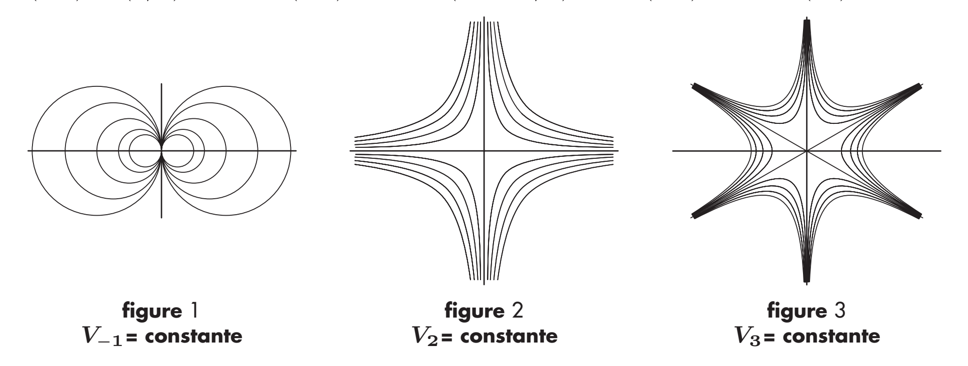


figure 1 — V_{-1} **= constante**

figure 2 — V_2 **= constante**

figure 3 — V_3 **= constante**

On peut interpréter V_2 et V_3 comme le potentiel-courant d'un écoulement plan non tourbillonnaire : les lignes de niveau sont les lignes de courant. La figure 2 illustre un courant entre deux cloisons rectilignes à angle droit (les axes) ; sur la figure 3, les cloisons, représentées par les asymptotes, forment des angles de 60 degrés.

Une propriété utile : le laplacien reste invariant par rotation des coordonnées.

Faisons en effet tourner les axes Ox, Oy d'un angle α. La nouvelle expression d'une fonction $f(r,\theta)$ est $g(r,\theta) = f(r,\theta+\alpha)$ et dans la formule du laplacien en coordonnées polaires, f et g ont même dérivée par rapport à θ. La propriété est vraie encore dans le cas de trois variables.

Pour des fonctions harmoniques en coordonnées cylindriques, voir l'exercice 9 ; en coordonnées sphériques, voir page 546.

4. Extremum local

Définition

Soit f une fonction de n variables définie sur un domaine $D \subset \mathbb{R}^n$ et soit A un point intérieur à D. On dit que f a un *maximum local* en A si l'on a $f(M) \leqslant f(A)$ pour tous les points $M \in D$ assez proches de A ; cela signifie :

il existe un nombre $r > 0$ tel qu'on a $f(M) \leqslant f(A)$ pour tout point M vérifiant $d(A, M) < r$.

On définit de même la notion de *minimum local*.

4.1 Conditions pour un extremum

Supposons par exemple que f est une fonction des deux variables x et y et que f a un maximum local en (a, b). Pour tout nombre x assez proche de a, on a $f(x, b) \leqslant f(a, b)$.

Ainsi la fonction d'une variable $g(x) = f(x, b)$ a un maximum local en a, donc sa dérivée $g'(a) = \dfrac{\partial f}{\partial x}(a, b)$ est nulle. De même, la fonction $h(y) = f(a, y)$ a un maximum local en b, donc $h'(b) = \dfrac{\partial f}{\partial y}(a, b) = 0$. Géométriquement, ces propriétés $\dfrac{\partial f}{\partial x}(a, b) = \dfrac{\partial f}{\partial y}(a, b) = 0$ signifient qu'au point $\big(a, b, f(a, b)\big)$ de la surface d'équation $z = f(x, y)$, le plan tangent est horizontal.

Condition nécessaire pour un extremum local. *Si une fonction de n variables a un extremum local en un point A intérieur à son domaine de définition, alors en ce point, toutes les dérivées partielles de f sont nulles.*

Définition

Un point où toutes les dérivées partielles de f sont nulles s'appelle un *point critique* de f.

Un extremum local de f est à chercher parmi les points critiques de f.

Mais, comme dans le cas d'une variable, la fonction ne présente pas forcément un extremum en un point critique. Par exemple, la fonction $(x,y) \mapsto x^3+y^3$ n'a pas d'extremum en $(0,0)$ bien que ses dérivées partielles à l'origine soient nulles (voir page 273).

Cherchons maintenant une condition permettant d'affirmer qu'il y a effectivement un maximum ou un minimum.

Supposons que $f(x,y)$ a des dérivées secondes continues et que le point $(0,0)$ est intérieur au domaine de définition. Sur la droite passant par l'origine et dirigée par un vecteur non nul (h,k), les valeurs de f sont données par $g(t) = f(th,tk)$. On a

$$\begin{cases} g'(t) = h\dfrac{\partial f}{\partial x}(th,tk)+k\dfrac{\partial f}{\partial x}(th,tk)\,, \text{ et} \\ g''(t) = h^2\dfrac{\partial^2 f}{\partial x^2}(th,tk)+2hk\dfrac{\partial^2 f}{\partial x \partial y}(th,tk)+k^2\dfrac{\partial^2 f}{\partial y^2}(th,tk) \end{cases}$$

Le développement limité à l'ordre 2 de g au point 0 s'écrit

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2}g''(0) + o(t^2)\,.$$

Supposons remplie la condition nécessaire pour que f ait un extremum local en $(0,0)$, c'est-à-dire $\dfrac{\partial f}{\partial x}(0,0) = \dfrac{\partial f}{\partial y}(0,0) = 0$. On a alors $g'(0) = 0$, donc

$$f(th,tk) - f(0,0) = g(t) - g(0) = \frac{t^2}{2}g''(0) + o(t^2)$$

On sait que pour t assez petit, cette différence a le signe de

$$g''(0) = h^2\frac{\partial^2 f}{\partial x^2}(0,0) + 2hk\frac{\partial^2 f}{\partial x \partial y}(0,0) + k^2\frac{\partial^2 f}{\partial y^2}(0,0).$$

Si par exemple $g''(0) > 0$, cela veut dire que l'on a $f(x,y) - f(0,0) > 0$ pour tout point (x,y) assez proche de l'origine dans la direction (h,k).

Condition suffisante pour un extremum local

Soit f une fonction de plusieurs variables ayant des dérivées secondes continues et soit A un point intérieur au domaine de définition.

Cas d'une fonction de deux variables. Supposons que f est une fonction de deux variables (x,y). Posons $p = \dfrac{\partial^2 f}{\partial x^2}(A)$, $q = \dfrac{\partial^2 f}{\partial x \partial y}(A)$ et $r = \dfrac{\partial^2 f}{\partial y^2}(A)$.

Proposition. *Supposons* $\dfrac{\partial f}{\partial x}(A) = \dfrac{\partial f}{\partial y}(A) = 0$ *et* $q^2 - rp < 0$.

- *Si $p > 0$ ou si $r > 0$, alors f a un minimum local en A.*
- *Si $p < 0$ ou si $r < 0$, alors f a un maximum local en A.*

Démonstration. Supposons pour simplifier $A = (0,0)$. D'après ce qui précède, pour que f ait un minimum local en A, il suffit que l'on ait $ph^2 + 2qhk + rk^2 > 0$ pour tout vecteur non nul (h,k). Lors de l'étude des produits scalaires dans $\mathbb{R}^2$, nous avons montré page 202 qu'il en est ainsi lorsque $p > 0$ et $q^2 - rp < 0$. Remarquons pour finir que si $q^2 - rp < 0$,

alors rp est positif, donc p et r sont de même signe. Supposons p et r négatifs. On a alors $-(ph^2+2qhk+rk^2) = (-p)h^2+2(-q)hk+(-r)k^2 > 0$, car $-p > 0$ et $(-q)^2 < (-r)(-p)$. Par suite, la fonction $-f$ a un minimum local en A, donc f a un maximum local en A. ■

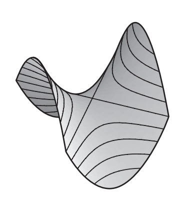
En posant $H = \begin{bmatrix} h \\ k \end{bmatrix}$ et $S = \begin{bmatrix} p & q \\ q & r \end{bmatrix}$, il vient $ph^2+2qhk+rk^2 = ({}^tH)SH$.

La proposition affirme que si la matrice symétrique S est définie positive, alors f a un minimum local en A.

Cas où $q^2 - rp > 0$. Le trinôme $pT^2+2qT+r$ ayant alors des racines réelles, il ne garde pas un signe constant et il en va de même de l'expression $ph^2+2qhk+rk^2 = k^2\big[p(h/k)^2+2q(h/k)+r\big]$.

Si $q^2 - rp > 0$, la fonction n'a pas d'extremum local en A.

Exemple 1. La fonction $f(x,y) = x^2 - y^2$ a un point critique en $(0,0)$ et l'on a ici $p=2$, $q=0$ et $r=-2$, donc $q^2-rp>0$.
En tout point $(x,0)$ différent de l'origine, on a $f(x,0)=x^2>0=f(0,0)$, et en tout point $(0,y)$ différent de l'origine, on a $f(0,y)=-y^2<0=f(0,0)$: il n'y a pas extremum local en $(0,0)$. Autour de ce point, la surface d'équation $z=x^2-y^2$ a la forme d'un col de montagne (voir page 21).



Exemple 2. Soit f une fonction harmonique sur un domaine D. En appelant comme précédemment p,q,r les dérivées secondes en un point intérieur M, on a $r=-p$, donc $q^2-rp = q^2+p^2$. Si p et q ne sont pas tous deux nuls en M, alors $q^2-rp>0$ et f n'a donc pas d'extremum local en M.
À moins d'être constante, une fonction harmonique n'a d'extremum local en aucun point intérieur de son domaine de définition : le maximum ou le minimum ne peut être atteint qu'en un point situé au bord du domaine. C'est ce qu'on voit sur les figures page 372.

Cas d'une fonction de n variables. Soit f une fonction de la variable vectorielle $X=(x_1,x_2,\dots,x_n)$ et soit A un point intérieur au domaine de définition. Donnons-nous un vecteur non nul $H=(h_1,h_2,\dots,h_n)$ et posons $g(t)=f(A+tH)$. On a $g'(t) = \sum_{i=0}^{n} h_i \dfrac{\partial f}{\partial x_i}(A+tH)$ et si l'on pose $a_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}(A)$, alors

$$g''(0) = \textstyle\sum_{i=0}^{n} a_{ii}h_i^2 + \sum_{0\leqslant i<j\leqslant n} 2a_{ij}h_ih_j \,.$$

Introduisons la matrice $S=[a_{ij}]$. C'est une matrice carrée de taille n qui est symétrique puisque $a_{ij}=a_{ji}$, d'après le théorème de Schwarz. Il vient alors $g''(0) = ({}^tH)SH$.

Définition
La matrice $\left[\dfrac{\partial^2 f}{\partial x_i \partial x_j}(A)\right]$ s'appelle la *matrice hessienne* de f au point A.

Le développement limité de g en 0 s'écrit $g(t) = g(0) + tg'(0) + \dfrac{t^2}{2} g''(0) + o(t^2)$. Supposons que la condition nécessaire d'extremum est vérifiée, c'est-à-dire que les dérivées partielles premières de f sont toutes nulles en A. Dans ce cas, $g'(0) = 0$ et $f(A+tH) - f(A) = g(t) - g(0) = \dfrac{t^2}{2}({}^tH)SH + o(t^2)$. Si la matrice S est définie positive, alors $({}^tH)SH > 0$, on a $f(A+tH) - f(A) > 0$ pour $t \neq 0$ assez petit et f a un minimum local en A.

Théorème. *Supposons que* $\dfrac{\partial f}{\partial x_i}(A) = 0$ *pour tout* i. *Si la matrice hessienne de* f *en* A *est définie positive, alors* f *a un minimum local en* A. *Si la matrice hessienne est définie négative,* f *a un maximum local en* A.

4.2 Méthode du gradient

Comme on ne sait pas, en général, résoudre analytiquement les équations permettant de trouver les points critiques, on utilise le plus souvent des méthodes numériques itératives pour trouver les extremums. La méthode du gradient est la plus simple. Supposons par exemple qu'on cherche un minimum local de f.

Principe de la méthode. On part d'un point X_0 et l'on calcule le gradient $\overrightarrow{\mathrm{Grad}}_f(X_0)$. Puisque le vecteur $-\overrightarrow{\mathrm{Grad}}_f(X_0)$ indique la direction des niveaux décroissants, on se déplace à partir de X_0 d'une longueur $h > 0$ dans la direction $-\overrightarrow{\mathrm{Grad}}_f(X_0)$, c'est-à-dire que l'on considère le point $X_0 - \dfrac{h}{\| \overrightarrow{\mathrm{Grad}}_f(X_0) \|} \overrightarrow{\mathrm{Grad}}_f(X_0)$. En pratique,

- on choisit un petit pas $h > 0$ et l'on progresse à pas de longueur h dans la direction $V_0 = -\dfrac{1}{\| \overrightarrow{\mathrm{Grad}}_f(X_0) \|} \overrightarrow{\mathrm{Grad}}_f(X_0)$ en formant les points

$$X_{0,1} = X_0 + hV_0 \quad , \quad X_{0,k+1} = X_{0,k} + hV_0 \ , \ \text{ pour } k = 1, 2, \dots$$

  tant que $f(X_{0,k+1}) < f(X_{0,k})$.

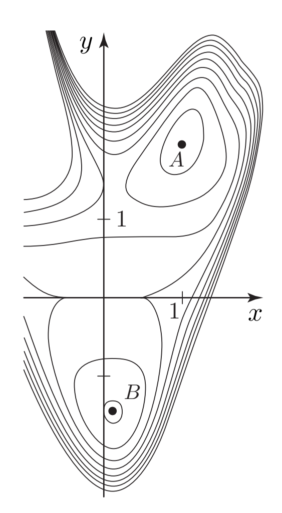
- Si X_1 est le dernier point $X_{0,N}$ obtenu, on recommence en prenant X_1 comme point initial. En itérant ce processus, on obtient des points $X_0, X_1, \dots$ de niveaux de plus en plus petits.

  Comme test d'arrêt, on peut prendre $\|\overrightarrow{\mathrm{Grad}}_f(X_n)\| < \varepsilon$, où ε est un petit nombre donné. Le calcul s'arrête alors en un point où les coordonnées du gradient sont inférieures à ε, mais il n'est pas facile de contrôler sérieusement la précision obtenue.

X_0 X_1 X_2 X_3

Pour la recherche d'un maximum, il faut bien sûr se déplacer dans la direction du gradient.

Exemple. La figure ci-contre montre des lignes de niveau de la fonction $f(x,y)=(x^2-y)^4+xy^3-4y^2$. Dans le domaine représenté, on voit qu'il y a deux extrema locaux en des points A et B situés aux environs de $(1\,,2)$ et de $(0{,}1\,,-1{,}4)$. En calculant quelques valeurs de f, on s'assure qu'il s'agit chaque fois d'un minimum local.



- Pour calculer les coordonnées de A, partons du point $X_0=(2{,}3)$. Avec $h=10^{-3}$ et $\varepsilon=10^{-1}$, on obtient en onze itérations le point de coordonnées 1,002, 2,004. Il est facile de vérifier que le minimum est en $A=(1{,}2)$, de valeur $f(A)=-7$.
- Cherchons les coordonnées de B en partant du point $X_0=(1,-2)$, avec $h=10^{-3}$ et $\varepsilon=10^{-2}$. En six itérations, on trouve le point $(0{,}1211\,,-1{,}4376)$, proche du minimum local $B=(0{,}12124\cdots\,,-1{,}43759\cdots)$ où la valeur de f est d'environ $-4{,}17$.

5. Extremum sous contraintes

Dans le paragraphe précédent, nos conditions pour un extremum local en A sont valables quand les variables sont libres de parcourir tout un voisinage de A. Mais souvent, on cherche à rendre maximum (ou minimum) une quantité $w=f(x,y)$ quand x et y sont liées par une condition $c(x,y)=k=$constante, c'est-à-dire contraintes de rester sur la ligne de niveau k de la fonction c.

Voici le théorème clef pour ce problème. Dans cet énoncé, g est une fonction de n variables ($n\geqslant 2$) à dérivées partielles continues et A est un point intérieur au domaine de définition de g.

Théorème des fonctions implicites. *Supposons que $g(A)=0$ et que $\dfrac{\partial g}{\partial x_n}(A)\neq 0$. Alors au voisinage de A, les solutions de l'équation $g(x_1,x_2,\ldots,x_n)=0$ sont données par $x_n=\varphi(x_1,\ldots,x_{n-1})$, où $\varphi(x_1,\ldots,x_{n-1})$ est une fonction à dérivées partielles continues. Si $A=(a_1,a_2,\ldots,a_n)$, alors $a_n=\varphi(a_1,\ldots,a_{n-1})$ et la fonction dérivée partielle $\dfrac{\partial\varphi}{\partial x_i}$ satisfait la relation $\dfrac{\partial g}{\partial x_i}(x_1,\ldots,x_n)+\dfrac{\partial g}{\partial x_n}(x_1,\ldots,x_n)\dfrac{\partial\varphi}{\partial x_i}(x_1,\ldots,x_{n-1})=0.$*

Dans la plupart des cas, il est impossible de résoudre explicitement l'équation $g(x_1,x_2,\ldots,x_n)=0$ en exprimant x_n en fonction de $x_1,\ldots,x_{n-1}$: le théorème affirme que, sous certaines conditions, cela est pourtant théoriquement possible. On dit que la fonction φ est *implicite*.

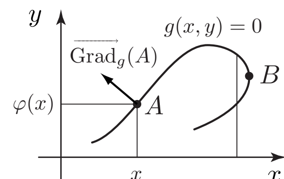
Pour tout $(x_1,x_2,\ldots,x_n)$ voisin de A, on a

$$g\big[x_1,x_2,\ldots,x_{n-1},\varphi(x_1,\ldots,x_{n-1})\big]=0$$

En dérivant cette identité, il vient $0 = \frac{\partial g}{\partial x_i}(X) + \frac{\partial g}{\partial x_n}(X)\frac{\partial \varphi}{\partial x_i}(x_1, \ldots, x_{n-1})$. Puisque $\frac{\partial g}{\partial x_n}(A) \neq 0$, cette égalité permet de calculer la i-ème dérivée partielle de φ en A.

La figure ci-dessous illustre le théorème dans le cas d'une fonction $g(x, y)$ de deux variables.

- C est la courbe d'équation $g(x, y) = 0$ et le point A est sur C.
- Puisque $\overrightarrow{\mathrm{Grad}}_g(A) = \left(\frac{\partial g}{\partial x}(A), \frac{\partial g}{\partial y}(A)\right)$, la condition $\frac{\partial g}{\partial y}(A) \neq 0$ signifie que le gradient en A n'est ni nul, ni horizontal.

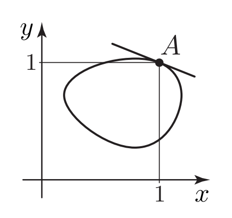


Comme on le voit sur la figure, si l'on reste au voisinage de A, alors pour x fixé, il y a sur la courbe exactement un point d'abscisse x. C'est donc qu'autour de A, la courbe C est le graphe d'une certaine fonction $\varphi(x)$.

Au point B au contraire, la tangente est verticale, donc $\frac{\partial g}{\partial y}(B) = 0$. Comme autour de B on peut trouver deux points de la courbe ayant même abscisse, C n'y est pas le graphe d'une fonction de x.

Cependant, autour de B, la courbe est le graphe d'une fonction de y, car $\frac{\partial g}{\partial x}(B) \neq 0$.

Exemple. Le graphe ci-contre représente la courbe d'équation $x^4+2y^4-2x-3y+2=0$. Elle passe au point $A = (1, 1)$ où l'on a $\frac{\partial g}{\partial y}(1, 1) = 5$. Il y a donc une fonction implicite φ telle que, pour x voisin de 1, la courbe ait pour équation $y = \varphi(x)$. La tangente en A a pour pente $\varphi'(1) = -\frac{\partial g}{\partial x}(1, 1)/\frac{\partial g}{\partial y}(1, 1) = -\frac{2}{5}$ et l'équation est $y-1 = -(2/5)(x-1)$.



Justification du théorème. Supposons $\frac{\partial g}{\partial y}(A) > 0$. Puisque la fonction $\frac{\partial g}{\partial y}$ est continue, on a encore $\frac{\partial g}{\partial y}(x, y) > 0$ pour (x, y) assez proche de A. Cela signifie que si x est assez proche de a, la fonction $y \mapsto g(x, y)$ est strictement croissante au voisinage de b. En appliquant cette propriété à la fonction $y \mapsto g(a, y)$ qui s'annule par hypothèse en $y = b$, on en déduit $g(a, y) < 0$ si $y < b$ et $g(a, y) > 0$ si $y > b$, du moins pour y assez proche de b. Choisissons de tels nombres $y_0 < b < y_1$. Comme g est continue, on a encore $g(x, y_0) < 0$ et $g(x, y_1) > 0$ pour x assez proche de a. Soit x une telle valeur. La fonction $y \mapsto g(x, y)$ étant continue et strictement croissante au voisinage de b, on en déduit, par le théorème des valeurs intermédiaires, que l'équation $g(x, y) = 0$ a une unique solution ; celle ci dépend de x : on la note $\varphi(x)$ et l'on montre que la fonction $x \mapsto \varphi(x)$ ainsi définie est dérivable. ■

5.1 Recherche d'un extremum sous contrainte

Extremum sous une seule contrainte

Le problème : rendre maximum (ou minimum) la quantité $w = f(x_1, x_2, \ldots, x_n)$ quand les variables sont contraintes de satisfaire la relation $g(x_1, x_2, \ldots, x_n) = 0$.

On suppose que f et g ont des dérivées partielles continues et que l'on a $\overrightarrow{\mathrm{Grad}}_g(X) \neq 0$ en tout point X satisfaisant la contrainte $g(X) = 0$.

Une condition nécessaire. *Pour que $w = f(X)$ soit maximum (ou minimum) en A sous la contrainte $g(X)=0$, il faut que les vecteurs $\overrightarrow{\mathrm{Grad}}_f(A)$ et $\overrightarrow{\mathrm{Grad}}_g(A)$ soient colinéaires.*

Démonstration. Supposons qu'il y a trois variables x, y, z et que le point $A = (a, b, c)$ est une solution du problème d'optimisation de $w = f(x, y, z)$ sous la contrainte $g(x, y, z) = 0$. Par hypothèse, le vecteur $\overrightarrow{\mathrm{Grad}}_g(A)$ est non nul : supposons par exemple $\dfrac{\partial g}{\partial z}(A) \neq 0$. D'après le théorème des fonctions implicites, la condition $g(x, y, z) = 0$ s'exprime au voisinage de A par $z = \varphi(x, y)$. Alors la fonction $(x, y) \mapsto f\big[x, y, \varphi(x, y)\big]$ a un extremum en (a, b), donc

$$(1) \qquad \frac{\partial f}{\partial x}(A) + \frac{\partial f}{\partial z}(A)\frac{\partial \varphi}{\partial x}(a, b) = 0 = \frac{\partial f}{\partial y}(A) + \frac{\partial f}{\partial z}(A)\frac{\partial \varphi}{\partial y}(a, b)$$

D'autre part, on sait que les dérivées partielles de φ en (a, b) vérifient

$$(2) \qquad \frac{\partial g}{\partial x}(A) + \frac{\partial g}{\partial z}(A)\frac{\partial \varphi}{\partial x}(a, b) = 0 = \frac{\partial g}{\partial y}(A) + \frac{\partial g}{\partial z}(A)\frac{\partial \varphi}{\partial y}(a, b)$$

De (1) et (2), on déduit qu'au point A, on a les deux premières relations suivantes, la troisième étant une identité :

$$\frac{\partial g}{\partial z}\frac{\partial f}{\partial x} - \frac{\partial f}{\partial z}\frac{\partial g}{\partial x} = 0$$
$$\frac{\partial g}{\partial z}\frac{\partial f}{\partial y} - \frac{\partial f}{\partial z}\frac{\partial g}{\partial y} = 0$$
$$\frac{\partial g}{\partial z}\frac{\partial f}{\partial z} - \frac{\partial f}{\partial z}\frac{\partial g}{\partial z} = 0$$

Ces égalités s'écrivent vectoriellement sous la forme $\dfrac{\partial g}{\partial z}(A)\,\overrightarrow{\mathrm{Grad}}_f\,(A) - \dfrac{\partial f}{\partial z}(A)\,\overrightarrow{\mathrm{Grad}}_g\,(A) = 0$.
Puisque $\dfrac{\partial g}{\partial z}(A) \neq 0$, les vecteurs $\overrightarrow{\mathrm{Grad}}_f(A)$ et $\overrightarrow{\mathrm{Grad}}_g(A)$ sont colinéaires. ■

Extremum à trois variables sous deux contraintes

Le problème : rendre maximum (ou minimum) la quantité $w = f(x, y, z)$ quand les variables sont contraintes de satisfaire les relations $g(x, y, z) = 0$ et $h(x, y, z) = 0$.
Supposons à nouveau que f, g et h ont des dérivées partielles continues et qu'en tout point X satisfaisant les contraintes $g(X) = 0 = h(X)$, les vecteurs $\overrightarrow{\mathrm{Grad}}_g(X)$ et $\overrightarrow{\mathrm{Grad}}_h(X)$ sont indépendants.

Une condition nécessaire. *Pour que $w = f(x, y, z)$ soit maximum (ou minimum) en un point A sous les contraintes $g(x, y, z) = h(x, y, z) = 0$, il faut que le déterminant des vecteurs $\overrightarrow{\mathrm{Grad}}_f(A)$, $\overrightarrow{\mathrm{Grad}}_g(A)$, $\overrightarrow{\mathrm{Grad}}_h(A)$ soit nul.*

Démonstration. Supposons que w est extremum en A et que l'indépendance des vecteurs gradients de g et de h en A est assurée parce que les deux dernières coordonnées de $\overrightarrow{\mathrm{Grad}}_g(A)$ et $\overrightarrow{\mathrm{Grad}}_h(A)$ ne sont pas proportionnelles : autrement dit, on suppose que le nombre $d = \dfrac{\partial h}{\partial y}(A)\dfrac{\partial g}{\partial z}(A) - \dfrac{\partial h}{\partial z}(A)\dfrac{\partial g}{\partial y}(A)$ n'est pas nul.

Si par exemple $\dfrac{\partial g}{\partial z}(A) \neq 0$, alors au voisinage de A, la condition $g(x,y,z)=0$ s'exprime sous la forme $z=\varphi(x,y)$. On en déduit que $u(x,y)=h\big[x,y,\varphi(x,y)\big]=0$, où la fonction φ vérifie $\dfrac{\partial g}{\partial z}\dfrac{\partial \varphi}{\partial y}=-\dfrac{\partial g}{\partial y}$. Au point A, on a $\dfrac{\partial g}{\partial z}\left(\dfrac{\partial h}{\partial y}+\dfrac{\partial h}{\partial z}\dfrac{\partial \varphi}{\partial y}\right)=d$. Le terme entre parenthèses est $\dfrac{\partial u}{\partial y}$ et puisque $d\neq 0$, on en déduit $\dfrac{\partial u}{\partial y}(A)\neq 0$. Au voisinage de A, la relation $u(x,y)=0$ s'exprime donc par $y=p(x)$. Finalement, les deux contraintes s'écrivent $y=p(x)$ et $z=\varphi\big(x,p(x)\big)=q(x)$. Pour tout x assez proche de l'abscisse a de A, on a ainsi identiquement

$$w=f\big[x,p(x),q(x)\big] \quad , \quad g\big[x,p(x),q(x)\big]=0 \quad \text{et} \quad h\big[x,p(x),q(x)\big]=0.$$

Quand x varie, w est extremum en a, donc la première fonction a une dérivée nulle en a ; les deux autres fonctions sont constantes (de valeur 0) au voisinage de a, donc leur dérivée en a est nulle également. Tout cela s'écrit :

$$\begin{aligned}
&\frac{\partial f}{\partial x}(A)+\frac{\partial f}{\partial y}(A)\,p'(a)+\frac{\partial f}{\partial z}(A)\,q'(a)=0\\
&\frac{\partial g}{\partial x}(A)+\frac{\partial g}{\partial y}(A)\,p'(a)+\frac{\partial g}{\partial z}(A)\,q'(a)=0\\
&\frac{\partial h}{\partial x}(A)+\frac{\partial h}{\partial y}(A)\,p'(a)+\frac{\partial h}{\partial z}(A)\,q'(a)=0
\end{aligned}$$

On voit que les colonnes de la matrice $\begin{bmatrix} \frac{\partial f}{\partial x}(A) & \frac{\partial f}{\partial y}(A) & \frac{\partial f}{\partial z}(A)\\ \frac{\partial g}{\partial x}(A) & \frac{\partial g}{\partial y}(A) & \frac{\partial g}{\partial z}(A)\\ \frac{\partial h}{\partial x}(A) & \frac{\partial h}{\partial y}(A) & \frac{\partial h}{\partial z}(A)\end{bmatrix} = \begin{bmatrix} {}^t\mathrm{Grad}_f(A)\\ {}^t\mathrm{Grad}_g(A)\\ {}^t\mathrm{Grad}_h(A)\end{bmatrix}$ ne sont pas indépendantes. Le déterminant des vecteurs gradients est donc nul. ■

Exemple : stratégie de consommation. Étant donné un panier de trois biens B_1, B_2 et B_3 en quantités x_1, x_2 et x_3, l'*utilité* du panier est $U(x_1,x_2,x_3)=x_1^a x_2^b x_3^c$. Ce nombre est un indice de satisfaction du consommateur après achat. Les exposants a,b,c sont des nombres positifs (choisis en général tels que $a+b+c=1$).

Si p_1, p_2 et p_3 sont les prix unitaires des biens, la dépense occasionnée par l'achat du panier est $p_1x_1+p_2x_2+p_3x_3$. Supposons qu'un consommateur dispose pour cet achat d'un revenu R. On constate qu'il va y dépenser tout son revenu en choisissant les quantités qui rendent maximum l'utilité. La détermination des quantités achetées se formule donc ainsi :

trouver x_1,x_2,x_3 *rendant maximum* $U(x_1,x_2,x_3)$ *quand* $p_1x_1+p_2x_2+p_3x_3=R$.

La contrainte sur les variables s'écrit

$$g(x_1,x_2,x_3)=0, \text{ où } g(x_1,x_2,x_3)=p_1x_1+p_2x_2+p_3x_3-R.$$

Le gradient de g est donc $\overrightarrow{\mathrm{Grad}}_g=(p_1,p_2,p_3)$. Calculons le gradient de U : on a

$$\frac{\partial U}{\partial x_1}=ax_1^{a-1}x_2^b x_3^c=\frac{a}{x_1}U(x_1,x_2,x_3)$$

et de même $\dfrac{\partial U}{\partial x_2} = \dfrac{b}{x_2} U(x_1, x_2, x_3)$, $\dfrac{\partial U}{\partial x_3} = \dfrac{c}{x_3} U(x_1, x_2, x_3)$, d'où

$$\overrightarrow{\mathrm{Grad}}_U(x_1, x_2, x_3) = U(x_1, x_2, x_3)\left(\frac{a}{x_1}, \frac{b}{x_2}, \frac{c}{x_3}\right)$$

Pour que les gradients de g et de u soient colinéaires, il doit exister un nombre λ tel que

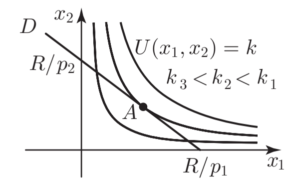
$$p_1 = \lambda \frac{a}{x_1} \quad , \quad p_2 = \lambda \frac{b}{x_2} \quad , \quad p_3 = \lambda \frac{c}{x_3}$$

Il vient alors $p_1 x_1 = \lambda a$, $p_2 x_2 = \lambda b$, $p_3 x_3 = \lambda c$; en ajoutant, on obtient $R = p_1 x_1 + p_2 x_2 + p_3 x_3 = \lambda(a+b+c)$ et $\lambda = \dfrac{R}{a+b+c}$. Finalement, la solution est

$$x_1 = \frac{aR}{p_1(a+b+c)} \quad , \quad x_2 = \frac{bR}{p_2(a+b+c)} \quad , \quad x_3 = \frac{cR}{p_3(a+b+c)}$$

Interprétation géométrique dans le cas de deux biens. Dans le cas de deux biens, la figure ci-contre montre la droite D d'équation $p_1 x_1 + p_2 x_2 = R$ et des lignes de niveau de la fonction $U(x_1, x_2) = x_1^a x_2^b$.

Si une courbe de niveau k coupe D en deux points, on voit qu'il y a sur D des points de niveau supérieurs à k : ces points d'intersection ne sont donc pas solution. Ainsi, la solution est géométriquement le point A de D où la ligne de niveau de U est tangente à D. En un point (x_1, x_2), le gradient de U est orthogonal à la ligne de niveau de U et le vecteur (p_1, p_2) est orthogonal à D. Le point A se caractérise donc bien par le fait que les vecteurs $\overrightarrow{\mathrm{Grad}}_U(A)$ et (p_1, p_2) sont colinéaires.



5.2 Une application statistique : le krigeage

Cette méthode d'estimation statistique, initiée par D.G. Krige, est très utilisée dans les Sciences de la Terre (Géologie, Océanographie) et dans la recherche minière. Nous en présentons la théorie, puis l'aspect pratique. Voici d'abord quelques définitions.

Espérance et Covariance. Si U est une variable aléatoire, rappelons que l'on note $E(U)$ l'espérance de U, si ce nombre existe (définition page 329). Soient U et V des variables aléatoires.

- Si U et V ont une espérance, alors $E(U+V) = E(U) + E(V)$. Il s'ensuit que la variable aléatoire $U - E(U)$ a une espérance nulle : on dit que U est *centrée*.
- L'espérance $E\big[\big(U - E(U)\big)\big(V - E(V)\big)\big]$ s'appelle la *covariance de U et V* et se note $\mathrm{Cov}(U, V)$. On a donc aussi $\mathrm{Cov}(U, V) = E(UV) - E(U)E(V)$.
- La *variance* de U est le nombre $\mathrm{Cov}(U, U) = E\Big[\big(U - E(U)\big)^2\Big] = E\big[U^2 - E(U)^2\big]$.

Pour les variables centrées $\overline{U} = U - E(U)$ et $\overline{V} = V - E(V)$, on a $\mathrm{Cov}(\overline{U}, \overline{V}) = \mathrm{Cov}(U, V)$, car $E(\overline{U}) = E(\overline{V}) = 0$.

Nous montrerons page 407 que si U et V sont indépendantes, alors $\mathrm{Cov}(U, V) = 0$; mais la réciproque n'est pas vraie.

Les données et le problème

Soit Z une fonction aléatoire, c'est-à-dire une famille de variables aléatoires $Z(x)$ dépendant d'un paramètre x. Faisons une première hypothèse.

Hypothèse (a) : Pour tous x,y, on a $E\big[Z(x)-Z(y)\big]=0$, autrement dit les variables aléatoires $Z(x)$ ont toutes la même espérance m.

Par définition de la covariance (supposée exister), on a alors pour tout x :

$$\text{(1)} \qquad \mathrm{Cov}(\big[Z(x), Z(x)\big] = E\big[Z(x)^2\big] - m^2$$

Donnons-nous n valeurs $x_1,x_2,\ldots,x_n$ du paramètre et les variables aléatoires $Z(x_1)$, $Z(x_2),\ldots,Z(x_n)$ correspondantes.

Le problème : À partir de cette donnée, trouver une estimation Z^* de la variable aléatoire $Z(x_0)$, où x_0 est une valeur du paramètre ne figurant pas parmi $x_1,x_2,\ldots,x_n$.

Exemple. Typiquement, $Z(x)$ est la teneur en minerai en un point x d'un gisement. On a prélevé des échantillons aux points $x_1,x_2,\ldots,x_n$ et mesuré des teneurs $z(x_1),z(x_2),\ldots,z(x_n)$. Le but est de faire la meilleure estimation statistique de la teneur $z(x_0)$ au point x_0. L'hypothèse (a) traduit une homogénéité dans les résultats de ces mesures, c'est-à-dire une certaine régularité géologique du bassin étudié.

Cherchons l'estimation Z^* sous la forme $Z^*=\sum_{k=1}^n \lambda_k Z(x_k)$, où les nombres λ_k vérifient $\sum_{k=1}^n \lambda_k = 1$; on définit ainsi une nouvelle variable aléatoire Z^*.

Justifions la condition sur les coefficients inconnus $\lambda_1,\ldots,\lambda_n$: puisque l'espérance d'une somme est la somme des espérances, l'espérance de $Z(x_0)-Z^*$ est $E\big[Z(x_0)-Z^*\big] = E\big[Z(x_0)\big] - \sum_{k=1}^n \lambda_k E\big[Z(x_k)\big] = m - \big(\sum_{k=1}^n \lambda_k\big)\, m$, car toutes les variables aléatoires $Z(x)$ ont même espérance m. Comme Z^* doit être une estimation de Z_0, cette espérance doit être nulle (on dit que l'estimation est sans biais) ; pour que cela soit vrai quel que soit m, il faut, d'après l'égalité ci-dessus, que la somme des λ_k soit égale à 1.

Si l'on peut choisir les coefficients λ_k de manière que $Z(x_0)-Z^*$ ait une variance v minimale, la variable aléatoire Z^* est une bonne estimation de $Z(x_0)$, par rapport à l'information connue. Puisque la somme des λ_k vaut 1, l'espérance de $Z(x_0)-Z^*$ est nulle, donc $v = E\Big[\big(Z(x_0)-Z^*\big)^2\Big]$. Le problème se formule ainsi :

trouver des coefficients λ_i rendant minimum la variance $v = E\Big[\big(Z(x_0)-Z^*\big)^2\Big]$.

Posons $\overline{Z}(x_0)=Z(x_0)-m$ et $\overline{Z}(x_k)=Z(x_k)-m$. Puisque $\sum_{k=1}^n \lambda_k=1$, on a $\sum_{k=1}^n \lambda_k \overline{Z}(x_k)= Z^*-m$, donc

$$Z(x_0)-Z^* = \big(Z(x_0)-m\big) - (Z^*-m) = \overline{Z}(x_0) - \sum_{k=1}^n \lambda_k \overline{Z}(x_k)$$

En développant le membre de droite, il vient

$$\big(Z(x_0)-Z^*\big)^2 = \overline{Z}(x_0)^2 - 2\sum_{k=1}^n \lambda_k \overline{Z}(x_0)\overline{Z}(x_k) + \sum_{i,j} \lambda_i\lambda_j \overline{Z}(x_i)\overline{Z}(x_j)$$

$$v = E\big[\overline{Z}(x_0)^2\big] - 2\sum_{k=1}^n \lambda_k E\big[\overline{Z}(x_0)\overline{Z}(x_k)\big] + \sum_{i,j} \lambda_i\lambda_j E\big[\overline{Z}(x_i)\overline{Z}(x_j)\big]$$

$$(2)\ \ v = \mathrm{Cov}\big[Z(x_0), Z(x_0)\big] - 2\sum_{k=1}^{n} \lambda_k \,\mathrm{Cov}\big[Z(x_0)Z(x_k)\big] + \sum_{i,j} \lambda_i\lambda_j \,\mathrm{Cov}\big[Z(x_i)Z(x_j)\big]$$

La variance v est fonction des n variables $(\lambda_1, \lambda_2, \dots, \lambda_n)$ liées par la condition $g(\lambda_1, \lambda_2, \dots, \lambda_n) = \lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$.

Pour que v soit minimum, il faut donc que les vecteurs $\overrightarrow{\mathrm{Grad}}_v = \left(\dfrac{\partial v}{\partial \lambda_1}, \dfrac{\partial v}{\partial \lambda_2}, \dots, \dfrac{\partial v}{\partial \lambda_n}\right)$ et $\overrightarrow{\mathrm{Grad}}_g = \left(\dfrac{\partial g}{\partial \lambda_1}, \dfrac{\partial g}{\partial \lambda_2}, \dots, \dfrac{\partial g}{\partial \lambda_2}\right) = (1,1,\dots,1)$ soient colinéaires. Calculons les dérivées partielles de v :

$$\frac{\partial v}{\partial \lambda_k} = -2\,\mathrm{Cov}\big[Z(x_0), Z(x_k)\big] + 2\sum_{j=1}^{n} \lambda_j \,\mathrm{Cov}\big[Z(x_k)Z(x_j)\big]$$

Les gradients sont colinéaires si et seulement s'il existe un nombre μ tel que

$$(3)\quad \sum_{j=1}^{n} \lambda_j \,\mathrm{Cov}\big[Z(x_k)Z(x_j)\big] - \mathrm{Cov}\big[Z(x_0), Z(x_k)\big] = \mu\ ,\ \text{ pour tout } k = 1, 2, \dots, n$$

Écrivons cela sous forme matricielle. En posant $c_{ij} = \mathrm{Cov}\big[Z(x_i)Z(x_j)\big]$, les égalités (3) et la relation $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$ se mettent sous la forme

$$(4)\qquad \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & & & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -\mu \end{bmatrix} = \begin{bmatrix} c_{01} \\ c_{02} \\ \vdots \\ c_{0n} \\ 1 \end{bmatrix}$$

La matrice ci-dessus est symétrique, de taille $n+1$. On voit que les coefficients inconnus $\lambda_1, \dots, \lambda_n$ et μ sont solutions d'un système linéaire. Cependant en pratique, les covariances c_{ij} ne sont aisées à déterminer directement. Pour continuer les calculs, faisons une hypothèse supplémentaire.

Hypothèse (b) : pour tout h, la covariance $\mathrm{Cov}\big[Z(x), Z(x+h)\big]$ ne dépend que de h (hypothèse de « stationnarité »). Posons $C(h) = \mathrm{Cov}\big[Z(x), Z(x+h)\big]$.

Exemple. Si $Z(x)$ est la teneur en minerai au point x d'un gisement, l'hypothèse (b) traduit le fait que l'influence exercée en un autre point y par la richesse en minerai en x ne dépend pas de la position de x, mais seulement du vecteur (ou de la distance) h représentant le déplacement entre x et y. Il faut définir convenablement le bassin pour pouvoir considérer que cette propriété est à peu près vérifiée. En général, la covariance $C(h)$ diminue quand $\|h\|$ augmente.

Comme conséquence de l'hypothèse, on trouve que la variance de $Z(x)$ est $C(0) = \mathrm{Cov}\big[Z(x), Z(x)\big]$, un nombre indépendant de x. L'égalité (1) s'écrit ainsi

$$(1')\qquad E\big[Z(x)^2\big] = C(0) + m^2$$

Le demi-variogramme. Avec la notation $\overline{Z}(x) = Z(x) - m$, on a $Z(x) - Z(+h) = \overline{Z}(x) - \overline{Z}(x+h)$; en développant le carré et en prenant l'espérance, on obtient d'après $(1')$

$$\begin{aligned} E\Big[\big(Z(x) - Z(x+h)\big)^2\Big] &= E\big[\overline{Z}(x)^2\big] + E\big[\overline{Z}(x+h)^2\big] - 2E\big[\overline{Z}(x)\overline{Z}(x+h)\big] \\ &= 2C(0) - 2C(h)\,. \end{aligned}$$

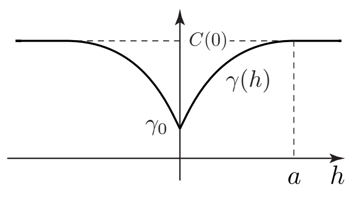
On définit le *demi-variogramme* en posant

$$\gamma(h) = \frac{1}{2}E\Big[\big(Z(x) - Z(x+h)\big)^2\Big] = C(0) - C(h), \quad \text{pour tout } h \neq 0.$$

On a $C(-h) = C(h)$ et donc $\gamma(-h) = \gamma(h)$.

En Géologie, la fonction γ permet de quantifier la notion de zone d'influence d'un échantillon : la covariance $C(h)$ diminue quand $\|h\|$ augmente, donc $\gamma(h)$ est fonction croissante de $\|h\|$: son taux de croissance mesure la diminution d'influence d'un échantillon de terrain sur des zones de plus en plus éloignées.

Quand $\|h\|$ est assez grand, il n'y a plus d'influence entre les points x et $x+h$ de sorte que la covariance $C(h)$ est presque nulle : on a $\gamma(h) \simeq C(0)$ pour $\|h\|$ suffisamment grand. Le demi-variogramme expérimental présente ainsi un palier horizontal à partir d'une valeur $\|h\| = a$ appelée portée.



un demi-variogramme expérimental

Quand h tend vers 0, on constate souvent expérimentalement que $\gamma(h)$ tend vers une valeur γ_0 strictement positive, et non pas vers 0 comme pourrait le laisser croire la définition théorique. En réalité, la décroissance de $\gamma(h)$ est très rapide quand $\|h\|$ est de l'ordre du diamètre moyen des grains du minéral considéré : à l'échelle des distances entre les sondages effectués, on observe une discontinuité en $h=0$ (effet « pépite »).

Les équations du krigeage

Pour $0 \leqslant i, j \leqslant n$, posons $\gamma_{ij} = \frac{1}{2}E\Big[\big(Z(x_i) - Z(x_j)\big)^2\Big] = C(0) - c_{ij}$.

$$\begin{aligned} \text{On a } \sum_{j=1}^{n} \gamma_{kj}\lambda_j &= \sum_{j=1}^{n} C(0)\lambda_j - \sum_{j=1}^{n} c_{kj}\lambda_j = C(0) - \sum_{j=1}^{n} c_{kj}\lambda_j \;, \text{ car } \sum_{j=1}^{n}\lambda_j = 1 \\ &= C(0) - c_{0k} - \mu = \gamma_{0k} - \mu \;, \quad \text{d'après (3)} \end{aligned}$$

Finalement, en posant

$$A = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & & & & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \;, \quad \Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} \quad \text{et} \quad B = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n} \\ 1 \end{bmatrix}$$

les coefficients $\lambda_1, \ldots, \lambda_n$ s'obtiennent en résolvant le système linéaire

(5) $$A\Lambda = B$$

Si l'on prend pour x_0 l'un des points x_p de $\{x_1, \ldots, x_n\}$, la solution Z^* est $Z(x_p)$: l'estimation est donc sans biais. Il reste à déterminer les coefficients du système.

Évaluation des coefficients de la matrice. Supposons que U est une variable aléatoire et que $(u_1, u_2, \ldots, u_N)$ est une réalisation de U, c'est-à-dire les résultats de N expériences. Si N est assez grand, la moyenne $\frac{1}{N}\sum_{i=1}^{N} u_i$, appelée *moyenne observée*, est une estimation statistique de l'espérance de U.

Dans notre cas, on ne dispose en général en chaque point x_k que d'une réalisation $z(x_k)$ (résultat de l'analyse du prélèvement en x_k). Mais considérons pour h donné, tous les couples de points x_p, x_q choisis parmi $x_1, \ldots, x_n$ et tels que $x_q - x_p = h$. Toutes les variables $\big(Z(x_p) - Z(x_q)\big)^2$ ont la même espérance $2\gamma(h)$, donc s'il y a $N(h)$ couples de points, on a $2\gamma(h) = \frac{1}{N(h)} \sum_{x_q - x_p = h} E\Big[\big(Z(x_p) - Z(x_q)\big)^2\Big]$. En prenant comme seules réalisations les $z(x_p)$ et si $N(h)$ est assez grand, on pourra néanmoins adopter la moyenne des $\big(z(x_q) - z(x_p)\big)^2$ comme valeur de $2\gamma(h)$.

C'est ainsi que l'on évalue les nombres γ_{ij}. Soient x_i et x_j deux des points x_1,x_2,…,x_n et soit $h = x_j - x_i$. Faisons la demi-moyenne des nombres $\big(z(x_p) - z(x_q)\big)^2$, où les points x_p et x_q sont pris parmi $x_1, \ldots, x_n$ et vérifient $x_q - x_p = h$, autrement dit, adoptons la *valeur empirique*

$$\gamma_{ij} = \frac{1}{2N(h)} \sum_{x_q - x_p = h} \big[z(x_q) - z(x_p)\big]^2$$

avec $h = x_j - x_i$, la sommation se faisant sur les $N(h)$ couples (x_p, x_q) tels que $x_q - x_p = h$.

Le demi-variogramme $\gamma(h)$ est alors connu pour les valeurs $h = x_j - x_i$, où $1 \leqslant i, j \leqslant n$.

Par exemple, si les points $x_1, \ldots, x_n$ sont alignés et régulièrement espacés (échantillonnage dans une veine de minerai), alors $\gamma_{12} = \gamma_{23} = \cdots = \gamma_{n-1\,n} = \frac{1}{2(n-1)} \sum_{p=1}^{n-1} \big(z(x_{p+1}) - z(x_p)\big)^2$.

x_1 x_2 x_3 x_4 x_5 x_6

Pour que la somme contienne un nombre suffisant de termes, on se donne en général une tolérance ε sur la valeur de h ; ainsi lorsque h est une distance, on fait intervenir, dans le calcul de la moyenne, les points tels que $h - \varepsilon \leqslant x_q - x_p \leqslant h + \varepsilon$.

Évaluation des coefficients du second membre. Puisqu'on cherche une estimation de $Z(x)$ en un point x_0 différent de $x_1, \ldots, x_n$, les nombres γ_{0k} ne font en général pas partie des valeurs $\gamma(h)$ précédemment déterminées. On effectue donc une interpolation à partir des points connus du demi-variogramme, par exemple au moyen d'une fonction spline (page 350). Cela fournit une formule analytique pour la fonction $\gamma(h)$, valable pour tout h, et pour γ_{0k}, on prend comme valeur $\gamma_{0k} = \gamma(x_0 - x_k)$.

En Géologie, on utilise souvent des formules adaptées à chaque type de demi-variogramme et dont l'efficacité est confirmée par l'expérience : cette interpolation constitue la partie délicate de la méthode.

Calcul final. Une fois déterminés tous les $\gamma_{i,j}$ pour $0 \leqslant i,j \leqslant n$, la résolution du système linéaire (5) donne les coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$ et l'estimation $z^* = \sum_{k=1}^n \lambda_k z(x_k)$ de $Z(x_0)$.
Calculons la variance minimale $v_{\min}$ ainsi déterminée. D'après les égalités (3), on a $\sum_{k,j} \lambda_k \lambda_j c_{kj} = \sum_{k=1}^n \lambda_k \mu + \sum_{k=1}^n \lambda_k c_{0k} = \mu + \sum_{k=1}^n \lambda_k c_{0k}$, puisque la somme des λ_k vaut 1. En reportant dans (2), il vient alors

$$v_{\min} = \operatorname{Cov}\big[Z(x_0), Z(x_0)\big] - 2\sum_{k=1}^n \lambda_k c_{0k} + \mu + \sum_{k=1}^n \lambda_k c_{0k} = C(0) - \sum_{k=1}^n \lambda_k c_{0k} + \mu$$

et en tenant compte de $c_{0k} = C(0) - \gamma_{0k}$, on obtient $v_{\min} = \sum_{k=1}^n \lambda_k \gamma_{0k} + \mu$.

Si Λ est le vecteur solution de (5), alors $v_{\min} = ({}^t\Lambda)B$.

Supposons par exemple que l'erreur d'estimation $Z^* - Z(x_0)$ est distribuée autour de la vraie valeur selon une loi normale. Alors la probabilité pour que $\big|Z^* - Z(x_0)\big| \leqslant \sqrt{v_{\min}}$ est d'environ 0,68 et la probabilité pour que $\big|Z^* - Z(x_0)\big| \leqslant 2\sqrt{v_{\min}}$ est d'environ 0,95 (voir page 335).

6. Intégrales à paramètre

Soit $f(t,x)$ une fonction à valeurs réelles. En l'intégrant par rapport à t sur un segment $[a,b]$, on obtient la fonction $F(x) = \int_a^b f(t,x)\,dt$ de la seule variable x. Nous allons voir que si f est assez régulière, alors la fonction F est dérivable et que $F'(x)$ se calcule en « dérivant sous le signe intégrale ».

Dérivation sous le signe intégrale. *Si les fonctions f et $\dfrac{\partial f}{\partial x}$ sont continues, alors*

$$\frac{d}{dx}\left(\int_a^b f(t,x)\,dt\right) = \int_a^b \frac{\partial f}{\partial x}(t,x)\,dt$$

La démonstration fait appel à une propriété des fonctions continues :

si $\varphi(t,h)$ est une fonction continue et si $\varphi(t,0) = 0$ pour tout $t \in [a,b]$, alors $\max_{t\in[a,b]} \big|\varphi(t,h)\big|$ tend vers 0 quand h tend vers 0.

Démonstration. Par définition de $\dfrac{\partial f}{\partial x}(t,x_0)$, on a

$$F(x_0+h) - F(x_0) = \int_a^b \big[f(t,x_0+h) - f(t,x_0)\big]\,dt = \int_a^b \Big[h\frac{\partial f}{\partial x}(t,x_0) + h\varphi(t,h)\Big]\,dt\,,$$

où $\varphi(t,h)$ est continue et $\varphi(t,0) = 0$. Il vient donc

$$F(x_0+h) - F(x_0) = h\int_a^b \frac{\partial f}{\partial x}(t,x_0)\,dt + h\int_a^b \varphi(t,h)\,dt$$

D'après les propriétés de l'intégrale, on a

$$\left|\int_a^b \varphi(t,h)\,dt\right| \leqslant \int_a^b \left|\varphi(t,h)\right| dt \leqslant |b-a|m(h), \quad \text{où } m(h) = \max_{t\in[a,b]} \left|\varphi(t,h)\right|$$

La propriété énoncée ci-dessus affirme que $\lim_{h\to 0} m(h) = 0$. Quand h tend vers 0, $h\int_a^b \varphi(t,h)\,dt$ est donc négligeable devant h. Par définition de la dérivée, cela veut dire que $\int_a^b \frac{\partial f}{\partial x}(t,x_0)\,dt$ est la dérivée de F en x_0. ■

Exemple d'optimisation en biométrie. Environ un quart d'heure après la pose d'une perfusion, la quantité d'une substance thérapeutique présente dans l'organisme varie en fonction du temps (compté en heures) selon la loi $q(t) = ae^{-0,2\,t} - be^{-1,8\,t}$, où a et b sont des paramètres dépendant des quantités initiales et sur lesquels on peut agir. L'efficacité maximale est obtenue lorsque la quantité de produit reste assez longtemps voisine d'une valeur optimale m. On mesure donc l'action sur une période de temps T par la fonction de contrôle $J = \int_0^T \left[q(t) - m\right]^2 dt$.

Le problème : trouver a et b pour que J soit minimal.

Calculons les dérivées partielles de J par rapport à a et à b en dérivant sous le signe intégrale. On a

$$\frac{\partial}{\partial a}\left[(q-m)^2\right] = 2\big(q(t)-m\big)\frac{\partial q}{\partial a} = 2\big(q(t)-m\big)e^{-0,2\,t}$$

$$\begin{aligned}\frac{\partial J}{\partial a} &= \int_0^T \frac{\partial}{\partial a}\left[\big(q(t)-m\big)^2\right] dt = 2\int_0^T (ae^{-0,4\,t} - be^{-2t} - me^{-0,2\,t})\,dt \\ &= 5(1-e^{-0,4\,T})a - (1-e^{-2T})b + 10(e^{-0,2T} - m)\end{aligned}$$

De même,

$$\begin{aligned}\frac{\partial J}{\partial b} &= \int_0^T \frac{\partial}{\partial b}\left[\big(q(t)-m\big)^2\right] dt = 2\int_0^T (-ae^{-2t} + be^{-3,6\,t} + me^{-1,8\,t})\,dt \\ &= -(1-e^{-2T})a + \frac{5}{9}(1-e^{-3,6\,T})b - \frac{10}{9}(e^{-1,8\,T} - m)\end{aligned}$$

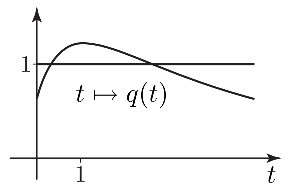
Pour que J soit minimum, il faut que ces deux dérivées partielles soient nulles. Supposons qu'on veuille exercer le contrôle sur cinq heures. On doit alors résoudre le système linéaire suivant aux inconnues a et b :

$$\begin{cases} 5(1-e^{-2})a - (1-e^{-10})b = 10(1-e^{-1})m \\ -9(1-e^{-10})a + 5(1-e^{-18})b = -10(1-e^{-9})m \end{cases}$$

La solution est environ $a = 1{,}71m$ et $b = 1{,}08m$. Calculons les dérivées secondes de J :

$$p = \frac{\partial^2 J}{\partial a^2} = 5(1-e^{-2})\ ,\ q = \frac{\partial^2 J}{\partial a \partial b} = -(1-e^{-10})5(1-e^{-2})\ ,\ r = \frac{\partial^2 J}{\partial b^2} = \frac{5}{9}(1-e^{-18})$$

On a $q^2 - rp < 0$ et p est positif, donc la fonction J atteint bien son minimum pour ces valeurs de a et b. Voici le graphe de la fonction $q(t)$ ainsi déterminée (on a pris $m=1$).



Sur la courbe, la partie décroissante qui intervient après environ une heure de traitement correspond à une phase d'élimination par l'organisme.

Énonçons pour finir les conditions qui permettent de dériver sous le signe intégrale généralisé.

Dérivation sous le signe intégrale généralisé. *Supposons que les fonctions* f *et* $\dfrac{\partial f}{\partial x}$ *sont continues et que l'intégrale généralisée* $\int_a^{+\infty} f(t,x)\,dt$ *existe. S'il y a une fonction* $g(t)$ *telle que* $\left|\dfrac{\partial f}{\partial x}(t,x)\right| \leqslant g(t)$ *pour tout* t *et tout* x*, et si l'intégrale généralisée* $\int_a^{+\infty} g(t)\,dt$ *existe, alors on a* $\dfrac{d}{dx}\left[\int_a^{+\infty} f(t,x)\,dt\right] = \int_a^{+\infty} \dfrac{\partial f}{\partial x}(t,x)\,dt$.

7. Linéarisation locale d'une transformation

Le problème. Considérons une transformation F de $\mathbb{R}^2$, c'est-à-dire une fonction $F : (x,y) \mapsto F(x,y)$ de deux variables, où $F(x,y) = \big(u(x,y), v(x,y)\big)$ est dans $\mathbb{R}^2$. Supposons que cette transformation a un point fixe X_0, autrement dit $F(X_0) = X_0$. On voudrait comprendre comment sont transformés les points voisins de X_0.
Pour simplifier, nous supposons que $X_0 = (0,0)$ (on se ramène à ce cas en considérant la transformation $X \mapsto F(X+X_0) - F(X_0)$).
Notre objectif est de trouver de nouvelles coordonnées dans lesquelles la transformation s'exprime plus simplement. Commençons par le cas où F est linéaire.

Cas d'une transformation linéaire. La transformation s'écrit $F(X) = AX$, où A est une matrice carrée de taille 2. Rappelons comment faire un bon changement linéaire de coordonnées (pages 186-188).
Supposons par exemple que la matrice A a deux valeurs propres réelles λ et μ distinctes. La matrice P des vecteurs propres étant inversible, nous pouvons faire le changement de coordonnées $X = PU$. Dans les nouvelles coordonnées $U = (u,v)$, la transformation $X' = AX$ s'écrit $U' = P^{-1}X' = P^{-1}AX = P^{-1}APU = DU$, où D est la matrice diagonale $\mathrm{diag}(\lambda,\mu)$. Les formules de transformation deviennent donc simplement $u' = \lambda u$, $v' = \mu v$.

Changement linéaire de coordonnées. Prenons une transformation $X'=F(X)$ dérivable telle que $F(0,0)=(0,0)$ et supposons que la matrice jacobienne $A=J_F(0,0)$ a deux valeurs propres réelles λ et μ distinctes. Faisons comme ci-dessus le changement de coordonnées $X = PU$, où P est la matrice des vecteurs propres de A. Dans les nouvelles coordonnées, la transformation s'écrit $U' = G(U)$, où

$$G(U) = U' = P^{-1}X' = P^{-1}F(X) = P^{-1}F(PU)$$

Soit $p:\mathbb{R}^2 \to \mathbb{R}^2$ l'application linéaire bijective définie par $p(U) = PU$. La matrice de la bijection réciproque p^{-1} est P^{-1}, donc $G(U) = (p^{-1} \circ F \circ p)(U)$. Si $P = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$, alors pour tout $U = (u,v)$, on a $p(U) = (\alpha u + \beta v, \gamma u + \delta v)$ et la matrice jacobienne de p est $J_p = P$. Puisque p^{-1} a pour matrice P^{-1}, on a de même $J_{p^{-1}} = P^{-1}$. La matrice jacobienne d'une composée étant le produit des matrices jacobiennes (page 367), il vient $J_G(0,0) = J_{p^{-1}}AJ_p = P^{-1}AP = \mathrm{diag}(\lambda,\mu)$, autrement dit

$$J_G(0,0) = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

$$\begin{array}{ccc} \mathbb{R}^2 & \xrightarrow{F} & \mathbb{R}^2 \\ p^{-1}\big| & & \big|p^{-1} \\ \mathbb{R}^2 & \xrightarrow[G]{} & \mathbb{R}^2 \end{array} \qquad \begin{array}{ccc} X & \xrightarrow{F} & X' \\ p^{-1}\big| & & \big|p^{-1} \\ U & \xrightarrow[G]{} & U' \end{array}$$

Par le changement de coordonnées $U = P^{-1}X$, on obtient une transformation G dont la matrice jacobienne au point fixe est diagonale.

Changement non linéaire de coordonnées. Reprenons la transformation $X'=F(X)$ et faisons un changement de coordonnées $U = \varphi(X)$ bijectif, où $X = (x,y)$ et $U = \big(u(x,y), v(x,y)\big)$. Supposons que :

i) $\varphi(0,0) = (0,0)$,

ii) φ et φ^{-1} ont des dérivées partielles continues,

iii) $J_\varphi(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$.

La dernière propriété signifie qu'au voisinage de l'origine, on a $u(x,y)=x+o\big(\|(x,y)\|\big)$ et $v(x,y) = y+o\big(\|(x,y)\|\big)$.

Puisque la composée $\varphi^{-1} \circ \varphi$ est l'identité, on a $J_{\varphi^{-1}}(0,0)J_\varphi(0,0) = I_2$, ou encore $J_{\varphi^{-1}}(0,0) = I_2$.
Dans les coordonnées (u,v), la transformation s'exprime par des formules $U' = G(U)$ telles que $U'=\varphi(X')=(\varphi\circ F)(X)=(\varphi\circ F)\big(\varphi^{-1}(U)\big)$, donc $G=\varphi\circ F\circ\varphi^{-1}$. Comme précédemment, on en déduit que la matrice jacobienne de G en $(0,0)$ est $J_G=J_\varphi J_F J_{\varphi^{-1}} = I_2 J_F I_2 = J_F$.

Ce changement de coordonnées ne modifie donc pas la matrice jacobienne en $(0,0)$. En particulier, les valeurs propres et les vecteurs propres sont inchangés.

Vers la linéarisation. Limitons-nous maintenant à une transformation F polynomiale : cela veut dire que $(x',y') = F(x,y)$ est de la forme

$$x' = ax+by+R_2(x,y)+R_3(x,y)+\cdots+R_n(x,y) \quad , \quad y' = cx+dy+S_2(x,y)+\cdots+S_n(x,y)$$

où $R_2(x,y) = r_{20}x^2 + r_{11}xy + r_{02}y^2$, $R_3(x,y) = r_{30}x^3 + r_{21}x^2y + r_{12}xy^2 + r_{03}y^3$, etc. Faisons d'abord, comme précédemment, un changement linéaire de coordonnées pour que la matrice jacobienne devienne diagonale, à coefficients les valeurs propres λ et μ de $J_F(0,0) = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Nous sommes maintenant ramenés à des formules pour x' et y' de la forme

(1) $x' = \lambda x + R_2(x,y) + R_3(x,y) + \cdots + R_n(x,y) \quad , \quad y' = \mu y + S_2(x,y) + \cdots + S_n(x,y)$

Nous allons chercher un changement de coordonnées $(u,v) = \varphi(x,y)$ polynomial ayant les propriétés indiquées au paragraphe précédent et tel que, dans les coordonnées (u,v), la transformation s'exprime par

$$(u',v') = (\lambda u, \mu v) + \textit{termes de degrés supérieurs à } N.$$

Si N est choisi assez grand, les termes de degré supérieurs à N seront très petits au voisinage de l'origine et la transformation $(u,v) \mapsto (u',v')$ sera ainsi très proche de la transformation linéaire $L(u,v) = (\lambda u, \mu v)$. De plus, nous savons que les valeurs propres et les vecteurs propres de L sont ceux de la transformation $X \mapsto J_F(0,0)X$, c'est-à-dire de l'approximation linéaire de F à l'origine.
Nous avons déjà fait un tel changement de coordonnées dans l'exemple 3 page 26.

Cherchons les fonctions u et v sous la forme :

$$u = x + U_2(x,y) + U_3(x,y) + \cdots + U_N(x,y) \quad , \quad v = x + V_2(x,y) + \cdots + V_N(x,y)$$

où $U_2 = u_{20}x^2 + u_{11}xy + u_{02}y^2$, $U_3(x,y) = u_{30}x^3 + u_{21}x^2y + u_{12}xy^2 + u_{03}y^3$, etc.
On a $u' = x' + U_2(x',y') + U_3(x',y') + \cdots + U_N(x',y')$, c'est-à-dire d'après (1) :

$$\begin{aligned} u' = {} & \big[\lambda x + R_2(x,y) + \cdots + R_n(x,y)\big] \\ & + U_2\big[\lambda x + R_2(x,y) + \cdots + R_n(x,y), \mu y + S_2(x,y) + \cdots + S_n(x,y)\big] \\ & + \cdots + U_N\big[\lambda x + R_2(x,y) + \cdots + R_n(x,y), \mu y + S_2(x,y) + \cdots + S_n(x,y)\big] \end{aligned}$$

et d'autre part

$$\lambda u = \lambda x + \lambda U_2(x,y) + \lambda U_3(x,y) + \cdots + \lambda U_N(x,y)\,.$$

Calculons les coefficients dans les U_i pour qu'il n'y ait plus de terme de degré 2 dans $u' - \lambda u$. Pour u', les termes de degré 2 sont localisés dans R_2 et U_2, et pour u, ils sont dans U_2.

► dans u', le terme en x^2 est $r_{20}x^2 + u_{20}\lambda^2 x^2$, donc on doit avoir

$$r_{20} + u_{20}\lambda^2 = \lambda u_{20}$$

► le terme en xy dans u' est $r_{11}xy + u_{11}\lambda\mu xy$, d'où

$$r_{11} + u_{11}\lambda\mu = \lambda u_{11}$$

► de même, en égalisant les termes en y^2 dans u' et λu, on obtient

$$r_{02} + u_{02}\mu^2 = \mu u_{02}$$

La première égalité s'écrit $(\lambda - \lambda^2)u_{20} = r_{20}$ et détermine u_{20} à condition que $\lambda \neq \lambda^2$.

Les deux autres égalités déterminent u_{11} et u_{02} à condition que $\lambda \neq \lambda\mu$ et $\mu \neq \mu^2$.
Les termes en x^3 dans u' sont les suivants :

$$\text{dans } R_3 : \left[r_{30}x^3\right] ; \qquad \text{dans } U_2 : \left[u_{20}(2\lambda x\, r_{20}x^2)\right] ; \qquad \text{dans } U_3 : \left[u_{30}\lambda^3 x^3\right]$$

On doit donc avoir l'égalité $r_{30} + 2r_{20}u_{20}\lambda + u_{30}\lambda^3 = \lambda u_{30}$. Puisque u_{20} est connu, cela détermine u_{30} si $\lambda \neq \lambda^3$.
Dans l'équation qui exprime l'égalité des termes en x^2y dans u' et λu, tous les termes sont connus, sauf u_{21} qui apparaît avec le coefficient $\lambda^2\mu$ dans u' et avec le coefficient λ dans λu ; on peut donc en tirer la valeur de u_{21} pourvu que $\lambda \neq \lambda^2\mu$. Les conditions pour trouver les coefficients de v sont analogues, mais les rôles de λ et μ sont échangés.
Nous n'avons pas vérifié que l'application $\varphi : (x,y) \mapsto (u,v)$ est bien un changement de coordonnées au voisinage de l'origine, c'est-à-dire ici une bijection dérivable dont la bijection réciproque est dérivable : c'est une conséquence du théorème des fonctions implicites.

En continuant ainsi par degré, on peut calculer de proche en proche les coefficients de u et de v qui annulent les termes de degré inférieur ou égal à N : la condition est qu'il n'existe entre les valeurs propres aucune relation de la forme

$$\lambda = \lambda^i\mu^j \text{ ou } \mu = \lambda^i\mu^j \ , \ \text{ avec } i \geqslant 0,\ j \geqslant 0 \text{ et } i+j \geqslant 2$$

On dit que ce sont les conditions de « non résonance ». En particulier, λ et μ doivent être non nuls, différents de ± 1 et l'on doit avoir $\lambda\mu \neq 1$ (sinon $\lambda = \lambda^2\mu$).
En négligeant les termes de degré supérieurs à N, on obtient dans les nouvelles coordonnées les formules $u' = \lambda u\ ,\ v' = \mu v$.

Si les conditions de non résonance sont vérifiées, il est même possible par cette méthode de trouver, au voisinage de l'origine, un changement de coordonnées qui supprime tous les termes de degré supérieur à 1, mais le changement de coordonnées n'est plus polynomial : dans ces nouvelles coordonnées, la transformation s'exprime alors exactement par les formules linéaires $u' = \lambda u, v' = \mu v$.

Exemple. Prenons la transformation $F(x,y) = \left(2x + x^2 - xy + y^3, (1/3)y + 2xy - y^2 + x^2y\right)$.
La matrice jacobienne à l'origine est $J_F = \begin{bmatrix} 2 & 0 \\ 0 & 1/3 \end{bmatrix}$. On voit facilement que les valeurs propres 2 et $1/3$ satisfont aux conditions de non résonance. Le changement de coordonnées polynomial qui supprime les termes de degré inférieurs ou égaux à 3 est donné par

$$u(x,y) = x - (1/2)x^2 - (3/4)xy + (1/3)x^3 - (15/8)x^2y + (63/64)xy^2 + (27/53)y^3$$
$$v(x,y) = y - 6xy - (9/2)y^2 + 25x^2y + 72xy^2 + (81/8)y^3$$

et l'on a $G(u,v) = (2u, v/3) +$ *termes de degré au moins* 4. Noter que, dans le changement de coordonnées, on ne peut pas trouver de formule explicite pour exprimer x et y au moyen de u, v.
Pour visualiser la transformation, nous avons représenté les itérés de quelques points proches de l'origine en reliant leurs positions successives par un segment ; dans la

figure 1, les coordonnées sont x, y ; dans la figure 2, les coordonnées sont u, v.

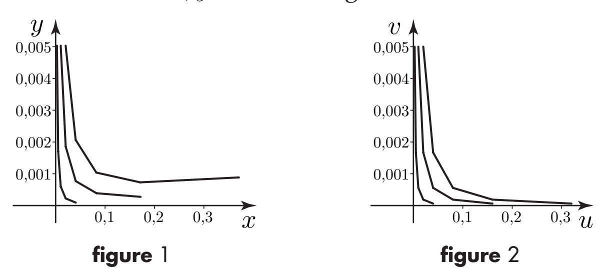


figure 1 **figure** 2

L'étude générale d'une telle itération linéaire est expliquée dans l'exemple 2 page 187 : comparer la figure 2 ci-dessus à celles qui y sont présentées.

Exercices

@ **1. Une recherche d'extremum.** Cherchons le maximum de la fonction $f(x,y) = x^3 - 3xy^2$ dans le disque unité $D = \{ (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leqslant 1 \}$.

a) Trouver les points critiques de f. Y a-t-il extremum local en ces points ?

b) Montrer que le maximum M et le minimum m de f sur D sont atteints en des points du cercle $x^2 + y^2 = 1$. Étudier la fonction $g(t) = f(\cos t, \sin t)$ pour trouver ces points et les valeurs M et m.

c) Montrer que la fonction f est harmonique.

2. En utilisant l'expression du laplacien en coordonnées polaires (page 372), vérifier que si k est un entier quelconque, les fonctions $r^k \cos k\theta$ et $r^k \sin k\theta$ sont harmoniques.

@ **3. Calcul d'une fonction d'offre.** Supposons que le coût de production à court terme d'une entreprise est $c(q,a) = q^3 - q^2 + (5-4a)q + 2a^2$, où q mesure le niveau de production et a l'impact des charges fixes.

a) Le coût à long terme d'une production q est la valeur minimum de $c(q,a)$ par rapport à a. En calculant $\dfrac{\partial c}{\partial a}$, montrer que le coût à long terme est $cl(q) = q^3 - 3q^2 + 5q$.

b) Soit p le prix de vente unitaire du bien fabriqué. Le profit d'une production q est donc $\Pi = pq - cl(q)$. Montrer que le profit positif maximum s'obtient pour $q = 1 + \sqrt{(p-2)/3}$ (vérifier que pour cette valeur, on a bien $\dfrac{\partial^2 \Pi}{\partial q^2} \leqslant 0$).

c) Le coût à long terme d'une unité de production est $cl(q)/q$. Pour quelle valeur de q ce coût est-il minimum ? Montrer que le minimum vaut $2{,}75$. En déduire que l'entreprise peut adopter le plan de production suivant : $q = \begin{cases} 0 & \text{si } p \leqslant 2{,}75 \\ 1 + \sqrt{(p-2)/3} & \text{si } p \geqslant 2{,}75 \end{cases}$.
Dessiner la courbe qui représente la variation de q en fonction de p lorsque p est entre 0 et 14.

4. Une stratégie de production. Supposons que la fonction de production d'une entreprise est de la forme $q = aT^\alpha K^\beta$, où T est la quantité de travail, K le capital mobilisé pour la production et α, β des nombres positifs tels que $\alpha + \beta = 1$, a étant un coefficient normatif fixe. On note w le salaire versé par unité de travail, r le taux de rémunération d'un capital immobilisé et f le montant des frais de gestion. Le coût de production est donc $c = wT + rK + f$.
L'entreprise assure une production Q donnée et veut minimiser son coût : comment doit-elle répartir quantité de travail et capital mobilisé ?

a) Calculer les vecteurs $\overrightarrow{\mathrm{Grad}}_q(T, K)$ et $\overrightarrow{\mathrm{Grad}}_c(T, K)$. Montrer que le minimum de c lorsque $q(T, K) = Q$ est obtenu lorsque $T/K = \alpha r/\beta w$.

b) En déduire que les valeurs cherchées sont $K = (Q/a)(\beta w/\alpha r)^\alpha$ et $T = (Q/a)(\alpha r/\beta w)^\beta$.

5. Une fonction $f(x, y)$ peut avoir une dérivée partielle $\dfrac{\partial f}{\partial y}$ identiquement nulle et dépendre néanmoins de y. Par exemple, définissons la fonction sur le domaine $D = \mathbb{R}^2 \setminus \{(x, 0) \mid x \geqslant 0\}$ en posant $f(x, y) = \begin{cases} x^2 & \text{si } x \geqslant 0 \text{ et } y > 0 \\ -x^2 & \text{si } x \geqslant 0 \text{ et } y < 0 \\ 0 & \text{si } x < 0 \end{cases}$.

a) Montrer que $\dfrac{\partial f}{\partial x}(a,b)$ vaut 0 si $a < 0$, $2a$ si a et b sont strictement positifs, et $-2a$ dans le cas $a > 0$, $b < 0$. Montrer que si $b \neq 0$, $\dfrac{f(h, b) - f(0, b)}{h}$ tend vers 0 quand h tend vers 0 (avec un signe quelconque) : en déduire que l'on a $\dfrac{\partial f}{\partial x}(0, b) = 0$ et que la fonction $\dfrac{\partial f}{\partial x}$ est continue en tout point de D.

b) Montrer que $\dfrac{\partial f}{\partial y} = 0$ en tout point de D, mais que les valeurs $f(1,y)$ dépendent de y.

@ **6.** Quelle est l'équation de la tangente à la courbe d'équation $x(x^2 + y^2) + y^3 = 8$ au point $(0, 2)$? Comment varie y quand x parcourt un petit intervalle autour de 0 ?

@ **7.** On s'intéresse à la distance de l'origine à la courbe (C) d'équation $x(x^2+y^2)+y^3=8$. Posons $f(x, y) = x(x^2 + y^2) + y^3 - 8$.

a) Soit M un point de (C) où la distance $d(M)$ à l'origine présente un extremum local. Montrer qu'il en va de même du carré de la distance à l'origine et que les coordonnées de M vérifient les équations $f(x, y) = 0$ et $y[x^2 + y^2 - 3xy] = 0$.

b) Pour tout point $M = (x, y)$ sur la courbe et tel que $M \neq (0, 2)$, on pose $y = tx$. Montrer que si $d(M)$ a un extremum local, alors t est l'un des trois nombres 0, $t_1 = (1/2)(3 + \sqrt{5})$ ou $t_2 = (1/2)(3 - \sqrt{5})$. Pour tout point de la courbe, calculer x et y au moyen de t. En déduire, par comparaison de trois distances, que le minimum de $d(M)$ ne peut être obtenu qu'au point $A = \big(2(1+t_1^2+t_1^3)^{-1/3}, 2t_1(1+t_1^2+t_1^3)^{-1/3}\big)$.

c) Dans la question précédente, on a obtenu un paramétrage de la courbe de la forme $x(t) = 2u(t), y(t) = 2tu(t)$, où $u(t) = (1+t^2+t^3)^{-1/3}$. En étudiant les variations de $d(t) = x(t)^2 + y(t)^2$ au voisinage de $t = t_1$, montrer que cette fonction y présente effectivement un minimum ; en déduire que la distance de l'origine à (C) est la distance OA. Montrer aussi qu'au point (2,0), la distance $d(M)$ a un minimum local.

d) Faire dessiner la courbe (C) par un ordinateur.

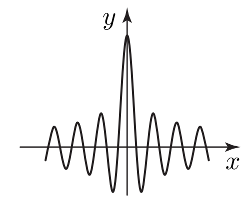
@ **8. Fonction de Bessel.** La *fonction de Bessel d'indice* 0 est $J_0(x) = \dfrac{1}{\pi}\displaystyle\int_0^\pi \cos(x\sin\theta)\,d\theta$.

a) Montrer que J_0 est une fonction paire. Calculer $J_0(0)$ et $J_0'(0)$.

b) Exprimer $J_0'(x)$ et $J_0''(x)$ au moyen d'une intégrale et montrer que $J_0''(x) + J_0(x) = \dfrac{1}{\pi}\displaystyle\int_0^\pi (\cos\theta)^2 \cos(x\sin\theta)\,d\theta$.

c) En faisant une intégration par parties dans l'expression de $J_0'(x)$, montrer que J_0 satisfait l'équation différentielle (b_0) : $y'' + \dfrac{1}{x}y' + y = 0$ *(équation de Bessel)*.

d) Vérifier au moyen d'un ordinateur que le graphe de J_0 a l'allure ci-dessous :

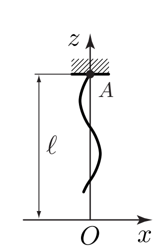


@ **9. Harmoniques cylindriques.** En coordonnées cylindriques, les fonctions harmoniques sont les solutions de l'équation $\dfrac{\partial^2 f}{\partial z^2} + \dfrac{\partial^2 f}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial f}{\partial r} + \dfrac{1}{r^2}\dfrac{\partial^2 f}{\partial \theta^2} = 0$ (page 372).
Une fonction dont les valeurs présentent une symétrie de révolution par rapport à l'axe Oz s'écrit $f(r,z)$ et, dans ce cas, l'équation devient $\dfrac{\partial^2 f}{\partial z^2} + \dfrac{\partial^2 f}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial f}{\partial r} = 0$.
Cherchons une solution de la forme $f(r,z) = e^{kz}u(r)$.

a) Montrer que la fonction f est harmonique à la condition que u soit solution de l'équation différentielle $u'' + \dfrac{1}{r}u' + k^2u = 0$.

b) En utilisant l'exercice **8**, montrer que pour toute constante k, les fonctions $e^{\pm k\,z}J_0(kr)$ sont harmoniques.

10. Petites oscillations d'une chaîne pesante. Considérons une chaîne pesante homogène de longueur ℓ suspendue à son extrémité supérieure A ; on note O la position d'équilibre de l'extrémité libre, on choisit un axe vertical Oz dirigé vers le haut et l'on repère le petit déplacement latéral d'un point de la chaîne sur un axe horizontal Ox. Les coordonnées de A sont donc $x = 0, z = \ell$. Si ρ est la densité linéaire de la chaîne, la tension en un point d'ordonnée z est $\rho g z$; en notant $x(z,t)$



le déplacement latéral à l'instant t du point de hauteur z et en laissant la chaîne osciller librement, l'équation du mouvement est

$$\rho \frac{\partial^2 x}{\partial t^2} - \frac{\partial}{\partial z}\left[\rho g z \frac{\partial x}{\partial z}\right] = 0 \tag{1}$$

a) Par le changement de variable $y = 2\sqrt{z/g}$, montrer que cette équation devient

$$\frac{\partial^2 x}{\partial t^2} - \left[\frac{\partial^2 x}{\partial y^2} + \frac{1}{y}\frac{\partial x}{\partial y}\right] = 0$$

b) Cherchons une solution de la forme $x = u(y)\cos \omega t$. Montrer que la fonction u doit satisfaire l'équation différentielle $u'' + \frac{1}{y}u' + \omega^2 u = 0$.

c) En utilisant l'exercice **8**, montrer que pour tout nombre a, la fonction $x(z,t) = aJ_0\left(2\omega\sqrt{z/g}\right)\cos\omega t$ est solution de (1). Décrire le mouvement de l'extrémité libre : que représente a ?

d) En exprimant que le point A est fixe, montrer que l'on a $J_0\left(2\omega\sqrt{\ell/g}\right) = 0$. En déduire que les pulsations ω possibles sont données par $\omega_i = (1/2)\alpha_i\sqrt{g/\ell}$, où $\alpha_1 < \alpha_2 < \cdots$ sont les solutions positives de l'équation $J_0(x) = 0$ (voir page **577**).

e) Notons $x_i(z,t)$ la solution correspondant à $\omega = \omega_i$. Montrer que dans ce mouvement, les points d'ordonnées $z_j = \ell(\alpha_j/\alpha_i)^2$, où $j = 1, 2, \ldots, i-1$, sont immobiles.

Si l'on se donne le profil de la chaîne à l'instant initial, le mouvement est une superposition de mouvements $x_i(z,t)$.

@11. Étude géométrique d'un ellipsoïde. On considère l'ellipsoïde E d'équation $f(x,y,z) = k$, où $f(x,y,z) = x^2 + 5y^2 + 6z^2 - 4xy - 8yz + 2xz - 2x + 8z + 6$, k étant un nombre strictement positif.

a) Calculer les dérivées partielles de f et trouver le point critique de f. Le centre de E est le point C où f atteint son minimum : en déduire que $C = (2, 0, -1)$.

b) On prend C comme nouvelle origine pour les coordonnées.

(i) Montrer que l'équation de E dans les nouvelles coordonnées X, Y, Z est $g(X,Y,Z) = k$, avec $g(X,Y,Z) = f(X+2, Y, Z-1)$.

(ii) Vérifier que les extrémités des axes de l'ellipsoïde sont les points M de E où la distance MC présente un maximum local (exercice **3** du chapitre **7**). Posons $w(X,Y,Z) = X^2 + Y^2 + Z^2 = MC^2$. En déduire qu'en ces points, les vecteurs $\overrightarrow{\mathrm{Grad}}_g(M)$ et $\overrightarrow{\mathrm{Grad}}_w(M)$ sont colinéaires. Calculer ces deux vecteurs gradients en un point M quelconque de coordonnées (X,Y,Z).

(iii) Chercher les extrémités en procédant comme dans l'exemple **3** page 219 : écrire la matrice symétrique S telle que $g(X,Y,Z) = [X\ Y\ Z]\, S \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$ et calculer son polynôme caractéristique ; montrer que les valeurs propres sont à peu près 10,012, 0,051 et 1,935 et en déduire des vecteurs propres approchés (rappelons que les

vecteurs propres de S sont deux à deux orthogonaux : voir page 217) ; déterminer enfin, pour chaque vecteur propre V, les deux nombres $\pm t$ tel que $g(tV) = k$ et calculer les coordonnées (x, y, z) des sommets de l'ellipsoïde.

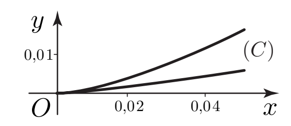
@12. Partie principale d'une quantité implicite. Supposons que des quantités x et y voisines de 0 sont reliées par la relation $f(x, y) = y^4 - 2x^3y^2 - 4x^5y + x^6 - x^7 = 0$.

a) Calculer les dérivées partielles de f en $(0, 0)$ et montrer que le théorème des fonctions implicites ne s'applique pas en ce point.

b) Pour x donné voisin de 0, l'équation $f(x,y)=0$ peut avoir plusieurs solutions $y(x)$. Pour trouver le comportement de ces solutions au voisinage de l'origine, cherchons une approximation $y(x) = ax^\alpha$, où $\alpha \neq 0$. Calculer $g(x) = f(x, ax^\alpha)$ et montrer que dans $g(x)$, les puissances de x ont pour exposant $4\alpha, 2\alpha + 3, \alpha + 5, 6$ et 7.

c) On cherche a et α pour que la plus petite puissance de x disparaisse. Il faut pour cela que le minimum ν des exposants soit obtenu au moins dans deux termes différents et qu'après regroupement, le coefficient de x^ν soit nul. En dessinant les graphes de $4\alpha, 2\alpha + 3, \alpha + 5, 6$ et 7 en fonction de α (ce sont des droites), montrer qu'on doit prendre $\alpha = 3/2$. Vérifier qu'on a alors $g(x) = x^6\left[(a^2-1)^2 - 4ax^{1/2} - x\right]$ et qu'il faut donc choisir $a = \pm 1$.

d) Dessiner le graphe de la fonction $y = x^{3/2}$ au voisinage de l'origine. La figure ci-contre montre la courbe (C) d'équation $f(x,y)=0$ pour $|x| \leqslant 0{,}05$: expliquez pourquoi ce dessin est compatible avec le résultat obtenu en **(c)** et pourquoi la courbe n'est pas représentée par le graphe d'une fonction de x.



e) Pour distinguer les deux branches de la courbe quand x est petit, cherchons une meilleure approximation sous la forme $y(x) = x^{3/2}(1 + bx^\beta)$. En posant $t^2 = x$, montrer que $f\big(x, y(x)\big) = t^{12}h(t)$, avec $h(t) = 4b^2t^{4\beta} + 4b^3t^{6\beta} + b^4t^{8\beta} - 4t - 4bt^{1+2\beta} - t^2$. Justifier comme précédemment que l'on a $\beta = 1/4$ et $b = \pm 1$. En déduire que les branches de la courbe sont approchées au voisinage de l'origine par les fonctions $y_1(x) = x^{3/2} + x^{7/4}$ et $y_2(x) = x^{3/2} - x^{7/4}$. Ces expressions s'appellent les *développements de Puiseux* de f au voisinage de l'origine.

On peut chercher un développement de Puiseux en un point où le théorème des fonctions implicites ne s'applique pas, c'est-à-dire où les deux dérivées partielles de la fonction sont nulles.

Chapitre 13

Intégrales multiples

1. Notion d'intégrale multiple et méthode de calcul

Dans ce chapitre, les coordonnées cartésiennes x, y, z des points de l'espace ou du plan sont prises dans un repère orthonormé.

Soit $f(x,y)$ une fonction continue de deux variables définie sur un rectangle $a \leqslant x \leqslant b$, $c \leqslant y \leqslant d$. En prenant une primitive en x, puis en y, on obtient une fonction F telle que $\dfrac{\partial}{\partial y}\left(\dfrac{\partial F}{\partial x}\right) = f$. Pour x fixé, si l'on intègre $f(x,y)$ par rapport à y entre c et d, on obtient une quantité

$$I(x) = \int_c^d f(x,y)\,dy = \int_c^d \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial x}\right)(x,y)\,dy = \frac{\partial F}{\partial x}(x,d) - \frac{\partial F}{\partial x}(x,c)$$

Intégrons maintenant $I(x)$ entre a et b :

$$(1)\quad \int_a^b I(x)\,dx = \int_a^b \frac{\partial F}{\partial x}(x,d)\,dx - \int_a^b \frac{\partial F}{\partial x}(x,c)\,dx = F(b,d) - F(a,d) - \big(F(b,c) - F(a,c)\big)$$

Intervertissons l'ordre des intégrations. Puisque $f = \dfrac{\partial^2 F}{\partial x \partial y}$ d'après le théorème de Schwarz, l'intégrale en x de a à b est

$$J(y) = \int_a^b f(x,y)\,dx = \int_a^b \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial y}\right)(x,y)\,dx = \frac{\partial F}{\partial y}(b,y) - \frac{\partial F}{\partial y}(a,y)\,,$$

et il vient

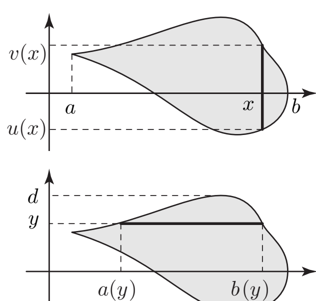
$$(2)\quad \int_c^d J(y)\,dy = \int_c^d \frac{\partial F}{\partial y}(b,y)\,dy - \int_c^d \frac{\partial F}{\partial y}(a,y)\,dy = F(b,d) - F(b,c) - \big(F(a,d) - F(a,c)\big)$$

Les nombres en (1) et (2) sont égaux. C'est donc que l'on a

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = \int_c^d \int_a^b f(x,y)\,dx\,dy\ ,$$

autrement dit on peut pratiquer la double intégration dans l'ordre qu'on veut. La quantité ainsi obtenue s'appelle l'intégrale double de f sur le rectangle $R=\{(x,y)\,|\,a\leqslant x \leqslant b, c \leqslant y \leqslant d\}$ et se note $\iint_R f(x,y)\,dxdy$.

Plus généralement, considérons un domaine D du plan, borné et limité par des courbes, comme sur les figures ci-contre :



- pour x fixé, y varie entre $u(x)$ et $v(x)$
- pour y fixé, x varie entre $a(y)$ et $b(y)$.

Alors les deux ordres d'intégration possibles conduisent encore au même résultat : si a et b sont les bornes extrêmes en x et si c et d sont les bornes extrêmes en y, on a l'égalité

$$\int_a^b \left(\int_{u(x)}^{v(x)} f(x,y)\,dy \right) dx = \int_c^d \left(\int_{a(y)}^{b(y)} f(x,y)\,dx \right) dy$$

Définition

La quantité ci-dessus s'appelle l'*intégrale double* de la fonction f sur le domaine D et se note $\iint_D f(x,y)\,dxdy$.

Une intégrale double se calcule par deux intégrations successives, dans l'ordre qu'on veut.

Exemple. Calculons l'intégrale double $I = \iint_D xy\,dxdy$, où D est formé des points à coordonnées positives dans le disque de rayon r centré à l'origine. On a donc $D = \{(x,y) \mid x > 0, y > 0,\ \text{et } x^2 + y^2 \leqslant r^2\}$. Pour x fixé entre 0 et r, la coordonnée y varie entre 0 et $\sqrt{r^2 - x^2}$, donc

$$I = \int_0^r \left(\int_0^{\sqrt{r^2-x^2}} xy\,dy \right) dx$$

On a

$$J_x = \int_0^{\sqrt{r^2-x^2}} xy\,dy = x \int_0^{\sqrt{r^2-x^2}} y\,dy = x \left[\frac{y^2}{2} \right]_0^{\sqrt{r^2-x^2}} = x\,\frac{r^2 - x^2}{2}$$

$$I = \int_0^r J_x\,dx = \frac{1}{2} \int_0^r x(r^2-x^2)\,dx = \frac{r^2}{2} \int_0^r x\,dx - \frac{1}{2} \int_0^r x^3\,dx = \frac{r^4}{4} - \frac{r^4}{8} = \frac{r^4}{8}.$$

Si $f(x,y,z)$ est une fonction de trois variables définie sur un domaine V de l'espace limité par des surfaces, on définit de même l'*intégrale triple* $\iiint_V f(x,y,z)\,dxdydz$: on la calcule en intégrant successivement par rapport à chaque variable, dans l'ordre qu'on veut.

Propriétés des intégrales multiples

a) Si f est à valeurs positives ou nulles, son intégrale est positive ou nulle.

b) Si f et g sont définies sur un même domaine D, l'intégrale de $f+g$ sur D est la somme des intégrales de f et de g sur D.

c) L'intégrale de f sur la réunion de deux domaines disjoints est la somme des intégrales de f sur ces domaines.

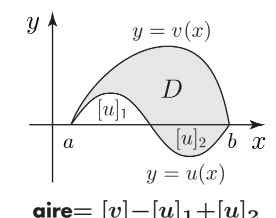
d) Si R est le rectangle $a \leqslant x \leqslant b$, $c \leqslant y \leqslant d$, alors pour un produit d'une fonction de x par une fonction de y, on a $\iint_R u(x)v(y)\,dxdy = \left(\int_a^b u(x)\,dx\right)\left(\int_c^d v(y)\,dy\right)$.

1.1 Aires et volumes

Calcul d'une aire plane

Soit D un domaine plan limité par des courbes, par exemple constitué des points (x,y) tel que $a \leqslant x \leqslant b$ et $u(x) \leqslant y \leqslant v(x)$. Pour x fixé entre a et b, on a $\int_{u(x)}^{v(x)} dy = v(x) - u(x)$, donc $\iint_D dxdy = \int_a^b \left[v(x) - u(x)\right] dx = \int_a^b v(x)\,dx - \int_a^b u(x)\,dx$.

L'intégrale $\int_a^b u(x)\,dx$ est l'aire algébrique comprise entre l'axe des abscisses et la courbe d'équation $y = u(x)$: sur la figure, c'est la quantité $[u]_1 - [u]_2$, où le crochet désigne l'aire absolue. De même, l'intégrale de v est l'aire $[v]$ sous cette courbe, donc la différence $\int_a^b v(x)\,dx - \int_a^b u(x)\,dx = [v] - [u]_1 + [u]_2$ est l'aire de D.



aire$= [v] - [u]_1 + [u]_2$

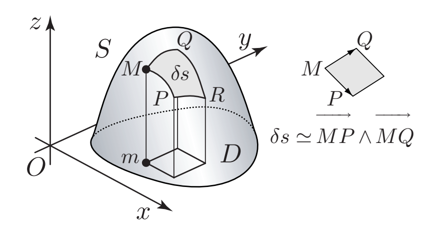
L'aire d'un domaine plan D est $\iint_D dxdy$.

Cette intégrale double s'interprète comme la somme des éléments d'aire infinitésimaux $dxdy$.

Calcul d'une aire dans l'espace

Soit S la surface d'équation $z = f(x,y)$ définie sur un domaine D du plan xOy.

Soient $m = (x,y)$ un point intérieur à D et $M = \big(x,y,f(x,y)\big)$ le point de S qui se projette en m. Donnons-nous de petits accroissements δx, δy, posons $x' = x + \delta x$, $y' = y + \delta y$ et considérons les points de la surface : $P = \big(x',y,f(x',y)\big)$, $Q = \big(x,y',f(x,y')\big)$ et $R = \big(x',y',f(x',y')\big)$. Supposons que f possède des dérivées partielles continues, donc S a un plan tangent en tout point. Si δx et δy sont très petits, l'aire δs du morceau de surface compris entre M, P, Q, R est proche de l'aire du parallélogramme construit sur les vecteurs $\overrightarrow{MP}, \overrightarrow{MQ}$ (c'est ainsi que nous avons déjà raisonné pour calculer la longueur d'une courbe paramétrée, page 312). L'aire du parallélogramme est la norme du produit vectoriel $\overrightarrow{MP} \wedge \overrightarrow{MQ}$ (page 225).



En ne gardant que les infiniments petits de l'ordre de δx et δy, on a

$$\overrightarrow{MP} = \left[\delta x\ ,\ 0\ ,\ f(x+\delta x, y) - f(x,y)\right] = \left[\delta x\ ,\ 0\ ,\ \frac{\partial f}{\partial x}\,\delta x\right]$$

$$\overrightarrow{MQ} = \left[0\ ,\ \delta y\ ,\ f(x, y+\delta y) - f(x,y)\right] = \left[0\ ,\ \delta y\ ,\ \frac{\partial f}{\partial y}\,\delta y\right]$$

$$\overrightarrow{MP} \wedge \overrightarrow{MQ} = \left[-\frac{\partial f}{\partial x}\delta x\delta y\ ,\ -\frac{\partial f}{\partial y}\delta x\delta y\ ,\ \delta x\delta y\right]$$

et pour l'élément d'aire sur la surface, il vient l'approximation

$$\delta s = \|\ \overrightarrow{MP} \wedge \overrightarrow{MQ}\ \| = |\delta x\delta y|\sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$$

En sommant tous ces éléments d'aire, on obtient la formule

$$\text{aire de } S\ = \iint_D \sqrt{1 + \left(\frac{\partial f}{\partial x}(x,y)\right)^2 + \left(\frac{\partial f}{\partial y}(x,y)\right)^2}\,dxdy$$

Cas d'une surface paramétrée. Supposons plus généralement que la surface S est l'ensemble des points

$$M(u,v) = \big(x(u,v), y(u,v), z(u,v)\big)$$

où u et v sont des paramètres réels, les fonctions $x(u,v), y(u,v), z(u,v)$ ayant des dérivées continues.

- Les vecteurs $\dfrac{\partial \overrightarrow{M}}{\partial u} = \left(\dfrac{\partial x}{\partial u}, \dfrac{\partial y}{\partial u}, \dfrac{\partial z}{\partial u}\right)$ et $\dfrac{\partial \overrightarrow{M}}{\partial v} = \left(\dfrac{\partial x}{\partial v}, \dfrac{\partial y}{\partial v}, \dfrac{\partial z}{\partial v}\right)$ sont tangents à S.
- Si en chaque point $M \in S$, ces vecteurs sont linéairement indépendants, on dit que S est *régulière*, ce qu'on suppose désormais.

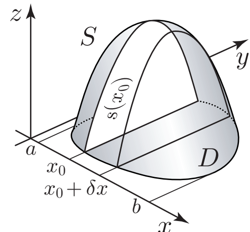
  Les vecteurs $\dfrac{\partial \overrightarrow{M}}{\partial u}(u,v), \dfrac{\partial \overrightarrow{M}}{\partial v}(u,v)$ *forment une base du plan tangent à* S *en* $M(u,v)$.
- Le produit vectoriel $\vec{H} = \dfrac{\partial \overrightarrow{M}}{\partial u} \wedge \dfrac{\partial \overrightarrow{M}}{\partial v}$ est un vecteur non nul et orthogonal à S. Le vecteur $\vec{N}(u,v) = \dfrac{1}{\|\ \vec{H}\ (u,v)\|}\ \vec{H}(u,v)$ s'appelle le *vecteur normal unitaire* au point $M(u,v)$.
- L'expression $d\mathrm{a} = \|\ \vec{H}\ (u,v)\|\ dudv$ s'appelle *l'élément d'aire de la surface.*

  C'est l'aire du parallélogramme tangent à S au point $M(u,v)$ et de cotés les « vecteurs tangents infiniment petits » $\dfrac{\partial \overrightarrow{M}}{\partial u}du,\ \dfrac{\partial \overrightarrow{M}}{\partial v}dv$.
- L'aire de S est $\int_S d\mathrm{a} = \iint_\Delta \|\ \vec{H}\ (u,v)\|\ dudv$, où Δ est le domaine parcouru par (u,v).

Une surface d'équation $z = f(x,y)$ est naturellement paramétrée en posant $x = u$, $y = v$, $z = f(u,v)$. On a alors $\dfrac{\partial \overrightarrow{M}}{\partial u} = \left(1, 0, \dfrac{\partial f}{\partial u}\right)$, $\dfrac{\partial \overrightarrow{M}}{\partial v} = \left(0, 1, \dfrac{\partial f}{\partial v}\right)$, $\vec{H} = \left(-\dfrac{\partial f}{\partial u}, -\dfrac{\partial f}{\partial v}, 1\right)$ et l'on retrouve la formule donnée précédemment pour l'aire.

Calculs de volumes

Une interprétation de l'intégrale double. Soit $f(x,y)$ une fonction à valeurs positives et soit S la surface d'équation $z=f(x,y)$, où (x,y) parcourt un domaine D du plan. Pour x_0 fixé, les points (x_0,y,z) situés sous la surface forment une portion de plan vertical limité par $c_0 \leqslant y \leqslant d_0$ et $0 \leqslant z \leqslant f(x_0,y)$. L'aire de cette portion de plan est donc $s(x_0) = \int_{c_0}^{d_0} f(x_0,y)\,dy$.



Épaississons la tranche en considérant les points (x,y,z) situés sous la surface et tels que $x_0 \leqslant x \leqslant x_0+\delta x$, où δx est une « quantité infiniment petite » (voir page 286).
En négligeant les produits de deux infiniment petits, le volume de cette tranche épaisse est le produit $s(x_0)\,\delta x$ de sa surface par son épaisseur. En sommant sur x_0, on obtient le volume V situé sous la surface.

Le volume compris entre un domaine plan D et la surface d'équation $z=f(x,y)$ est

$$\iint_D f(x,y)\,dxdy\,.$$

Une interprétation de l'intégrale triple. Soit V un domaine de l'espace compris entre deux surfaces d'équation $z=u(x,y)$ et $z=v(x,y)$, où $u(x,y) \leqslant v(x,y)$, les coordonnées x et y parcourant un domaine D du plan. On a ainsi $V = \{(x,y,z) \mid (x,y) \in D \text{ et } u(x,y) \leqslant z \leqslant v(x,y)\}$. Si (x,y) est un point fixé, alors $\int_{u(x,y)}^{v(x,y)} dz = v(x,y)-u(x,y)$, donc

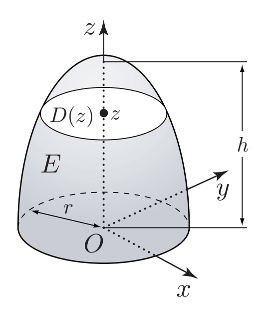
$$\iiint_V dxdydz = \iint_D \big[v(x,y)-u(x,y)\big]\,dxdy = \iint_D v(x,y)\,dxdy - \iint_D u(x,y)\big)\,dxdy$$

L'intégrale double $\iint_D v(x,y)dxdy$ est le volume compris entre le plan xOy et la surface $z=v(x,y)$, l'intégrale double $\iint_D u(x,y)\,dxdy$ est celui compris entre le plan xOy et la surface $z=u(x,y)$, donc le volume de V est $\iiint_V dxdydz$.

Le volume d'une portion V d'espace est $\displaystyle\iiint_V dxdydz$.

L'intégrale s'interprète comme la somme des éléments de volume infinitésimaux $dxdydz$.

Exemple. Calculons le volume de l'ellipsoïde (plein) de révolution d'axe Oz, de rayon r dans le plan xOy et de hauteur $2h$. Son équation est $x^2+y^2+\left(\frac{rz}{h}\right)^2 = r^2$. Notons E l'ellipsoïde. Pour calculer son volume $V = \iiint_E dxdydz$, intégrons d'abord par rapport à x et y, c'est-à-dire pour z fixé. À une hauteur z comprise entre $-h$ et h, la section horizontale est



un disque $D(z)$ limité par le cercle d'équation $x^2+y^2=r^2\left(1-\frac{z^2}{h^2}\right)$; l'aire du disque $D(z)$ est $s(z)=\pi r^2\left(1-\frac{z^2}{h^2}\right)$ et $V=\int_{-h}^{h}\left(\iint_{D(z)} dxdy\right)\ dz$, donc

$$V=\int_{-h}^{h} s(z)\,dz=\pi r^2\int_{-h}^{h}\left(1-\frac{z^2}{h^2}\right)dz=\pi r^2\left(2h-\frac{2h^3}{3h^2}\right)=\frac{4}{3}\pi r^2 h$$

En particulier, pour $h=r$, on trouve que le volume d'une boule de rayon r est $(4/3)\pi r^3$.

Centre de gravité, moment d'inertie

Choisissons un point O de l'espace. Si l'on place en des points $A_1,\ldots,A_n$ des masses $m_1,\ldots,m_n$, le centre de gravité du système ainsi formé est le point G (indépendant du choix de l'origine O) défini par

$$\overrightarrow{OG}=\frac{1}{M}\sum_{i=1}^{n} m_i\overrightarrow{OA_i}\,,$$

où $M=\sum_{i=1}^{n} m_i$ est la masse totale du système. En prenant des axes passant par O, les points A_i ont des coordonnées (x_i,y_i,z_i) et les coordonnées de G sont

$$\frac{1}{M}\left(\sum_{i=1}^{n} m_i x_i\,,\ \sum_{i=1}^{n} m_i y_i\,,\sum_{i=1}^{n} m_i z_i\right)$$

On définit de même le *centre de gravité d'un solide* S : la masse d'un élément de volume infinitésimal placé en (x,y,z) est $dm=\rho(x,y,z)\,dxdydz$, où ρ est la masse volumique en chaque point du solide.

► La masse totale du solide est $M=\iiint_S \rho(x,y,z)\,dxdydz$.

► Les coordonnées du centre de gravité sont

$$\frac{1}{M}\iiint_S x\,dm\ ,\ \ \frac{1}{M}\iiint_S y\,dm\ ,\ \ \frac{1}{M}\iiint_S z\,dm$$

Le *moment d'inertie* du solide par rapport à un axe Δ est $\iiint_S (d_M)^2\rho\,dxdydz$, où d_M est la distance du point $M=(x,y,z)$ à la droite Δ et ρ la masse volumique en ce point. Nous avons aussi défini page 227 la matrice d'inertie relativement à un point.

Exemple 1. Reprenons l'ellipsoïde de l'exemple précédent et considérons une calotte supérieure, pleine, de hauteur a, où $0<a\leqslant h$. Cette calotte étant située entre les hauteurs $h-a$ et h, son volume est

$$\begin{aligned} v&=\int_{h-a}^{h}\pi r^2\left(1-\frac{z^2}{h^2}\right)dz=\pi r^2\left[a-\frac{h^3-(h-a)^3}{3h^2}\right]\\ &=\pi r^2 h\left[1-k-\frac{1-k^3}{3}\right]\ ,\ \text{où } k=\frac{h-a}{h}. \end{aligned}$$

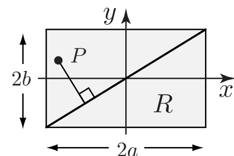
Supposons que la calotte (C) est un solide homogène de masse volumique ρ. Le centre de gravité G de la calotte se trouve sur l'axe de symétrie Oz : les coordonnées

de G sont $(0,0,Z)$, où $Z = \dfrac{1}{\rho v}\iiint_C z\,\rho dxdydz$. Il vient

$$vZ = \int_{h-a}^{h} z\,\pi r^2\left(1-\frac{z^2}{h^2}\right)dz = \pi r^2\int_{h-a}^{h}\left(z-\frac{z^3}{h^2}\right)dz$$
$$= \pi r^2\left[\frac{h^2-(h-a)^2}{2}-\frac{h^4-(h-a)^4}{4h^2}\right] = \pi r^2 h^2\left[\frac{1-k^2}{2}-\frac{1-k^4}{4}\right]$$

d'où l'on tire la valeur de Z. Si $a=h=r$, notre calotte pleine est la demi-boule supérieure de rayon r centrée à l'origine : le centre de gravité est à la hauteur $Z=(3/8)r$.

Exemple 2. Considérons une plaque rectangulaire homogène, de longueur $2a$, de largeur $2b$ et de masse M. Quel est le moment d'inertie de la plaque par rapport à l'une de ses diagonales ?



Prenons l'origine du repère au centre du rectangle et l'axe des abscisses dans le sens de la longueur. Le domaine R est donc formé des points (x,y) tels que $-a \leqslant x \leqslant a$ et $-b \leqslant y \leqslant b$. La diagonale de pente positive a pour équation $y=(b/a)x$, ou encore $bx-ay=0$. La distance d'un point $P=(x,y)$ à cette droite est $d(x,y)=\dfrac{|bx-ay|}{\sqrt{a^2+b^2}}$. La masse surfacique, rapport de la masse à l'aire, est $\rho=M/4ab$. Le moment d'inertie de la plaque par rapport à une diagonale est donc par définition

$$I = \rho\iint_R d(x,y)^2\,dxdy = \frac{M}{4ab(a^2+b^2)}\int_{-a}^{a}\left(\int_{-b}^{b}(bx-ay)^2\,dy\right)dx$$

On a

$$\int_{-b}^{b}(bx-ay)^2\,dy = \frac{-1}{3a}\Big[(bx-ay)^3\Big]_{-b}^{b} = \frac{1}{3a}\left[(bx+ab)^3-(bx-ab)^3\right] = \frac{b^3}{3a}\left[(x+a)^3-(x-a)^3\right]$$

$$\int_{-a}^{a}\left[(x+a)^3-(x-a)^3\right]dx = \frac{(2a)^4}{4}+\frac{(2a)^4}{4} = 8a^4$$

d'où $I = \dfrac{M}{4ab(a^2+b^2)}\dfrac{b^3}{3a}8a^4 = \dfrac{M}{4ab(a^2+b^2)}\dfrac{8a^3b^3}{3} = \dfrac{2M}{3}\dfrac{a^2b^2}{a^2+b^2}$.

1.2 Changement de variables

Soit D un domaine de l'espace dont les points (x,y,z) sont paramétrés par trois variables (u,v,w) : précisément, supposons qu'il y a une application $\varphi:\Delta\to D$ de la forme $(u,v,w)\mapsto\big(x(u,v,w),y(u,v,w),z(u,v,w)\big)$ ayant les propriétés suivantes :

- $\varphi:\Delta\to D$ est une bijection,
- les fonctions x, y et z ont des dérivées partielles continues par rapport à u,v,w,
- en tout point $(u,v,w)\in\Delta$, la matrice jacobienne J_φ est inversible (page 367).

La formule du changement de variable pour l'intégrale ordinaire (page 320) se généralise aux intégrales multiples, sous la forme :

$$(*)\qquad \iiint_D f(x,y,z)\,dxdydz = \iiint_\Delta f\big(\varphi(u,v,w)\big)\,\big|\det J_\varphi\big|\,dudvdw$$

Expliquons pourquoi intervient le déterminant de la matrice jacobienne. Considérons dans Δ un petit parallélépipède P de sommet $U=(u,v,w)$, à côtés parallèles aux axes et de longueurs $\delta u, \delta v, \delta w$. Les côtés de P sont donc dirigés par les vecteurs $(\delta u, 0, 0)$, $(0, \delta v, 0)$, $(0, 0, \delta w)$. Posons $M=\varphi(U)$. Par l'approximation affine de φ en U, un point quelconque U' de P se transforme en $M' = M + J_\varphi X$, où X est la colonne des coordonnées de $\overrightarrow{UU'}$. En appelant J_1, J_2, J_3 les colonnes de la matrice jacobienne, l'image de P est le parallélépipède Π de sommet M construit sur les vecteurs

$$J_\varphi(U)\begin{bmatrix}\delta u\\0\\0\end{bmatrix}=(\delta u)J_1 \quad , \quad J_\varphi(U)\begin{bmatrix}0\\\delta v\\0\end{bmatrix}=(\delta v)J_2 \quad , \quad J_\varphi(U)\begin{bmatrix}0\\0\\\delta w\end{bmatrix}=(\delta w)J_3$$

Puisque le volume de Π est la valeur absolue du déterminant de ces vecteurs, il vient

$$\text{volume de } \Pi = \left|\det(J_1, J_2, J_3)\right|\delta u\delta v\delta w = \left|\det\big(J_\varphi(U)\big)\right|\delta u\delta v\delta w$$

En considérant que P et Π sont des « éléments de volume infiniment petits », l'intégrale triple $\iiint_D f(x,y,z)\,dxdydz$ s'obtient en sommant toutes les quantités

$$f(M)\times \text{volume de } \Pi = f\big(\varphi(U)\big)\left|\det\big(J_\varphi(U)\big)\right|\delta u\delta v\delta w\,.$$

Intégrale double en coordonnées polaires. Rappelons que si $M=(x,y)$ est un point du plan différent de l'origine O, les coordonnées polaires de M sont les nombres $r = OM = \sqrt{x^2+y^2}$ et $\theta = \widehat{\overrightarrow{Ox}, \overrightarrow{OM}}$. On a donc les formules

$$x = r\cos\theta \quad , \quad y = r\sin\theta \ , \ \text{ où } \ r>0 \text{ et } 0 \leqslant \theta < 2\pi.$$

Le changement de variable est $\varphi(r,\theta)=(r\cos\theta, r\sin\theta)$ et le déterminant de la matrice jacobienne (page 368) est $\det J_\varphi = \begin{vmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{vmatrix} = r(\cos\theta)^2 + r(\sin\theta)^2 = r.$

Si D est un domaine plan et si Δ est l'ensemble des coordonnées (r,θ) des points de D, la formule du changement de variables s'écrit

$$\iint_D f(x,y)\,dxdy = \iint_\Delta f(r\cos\theta, r\sin\theta)\,r\,drd\theta$$

Exemple. Sur la sphère de rayon R centrée à l'origine, découpons la calotte supérieure de hauteur $h \leqslant R$. Son plan de base est à la distance $d = R-h$ de l'origine, le cercle horizontal inférieur a pour rayon $\rho = \sqrt{R^2-d^2}$, donc la calotte est la surface d'équation $z=\sqrt{R^2-x^2-y^2}$, où $x^2+y^2 \leqslant \rho^2$. Calculons l'aire A de cette calotte en utilisant la formule page 400. Les dérivées partielles de la fonction $f(x,y)=\sqrt{R^2-x^2-y^2}$ sont $\dfrac{\partial f}{\partial x} = \dfrac{-x}{\sqrt{R^2-x^2-y^2}}$, $\dfrac{\partial f}{\partial y} = \dfrac{-y}{\sqrt{R^2-x^2-y^2}}$, donc $1+\left(\dfrac{\partial f}{\partial x}\right)^2+\left(\dfrac{\partial f}{\partial y}\right)^2 = \dfrac{R^2}{R^2-x^2-y^2}$. En appelant D le disque du plan xOy défini par $x^2+y^2 \leqslant \rho^2$, il vient

$$A = \iint_D \sqrt{1+\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2}\,dxdy = R\iint_D \frac{1}{\sqrt{R^2-x^2-y^2}}\,dxdy$$

En coordonnées polaires, le disque D est défini par $r \leqslant \rho$ et $0 \leqslant \theta < 2\pi$. On a donc

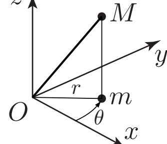
$$A = R\int_0^{2\pi}\left(\int_0^{\rho}\frac{1}{\sqrt{R^2-r^2}}r\,dr\right)d\theta = 2\pi R\int_0^{\rho}\frac{r}{\sqrt{R^2-r^2}}\,dr$$

Par le changement de variable $u = \sqrt{R^2-r^2}$, on a $du = \dfrac{-r\,dr}{\sqrt{R^2-r^2}}$, d'où

$$A = 2\pi R\left[-\sqrt{R^2-r^2}\right]_{r=0}^{r=\rho} = 2\pi R\big(R-\sqrt{R^2-\rho^2}\big) = 2\pi R(R-d) = 2\pi Rh$$

En particulier, pour $h = R$, on trouve que l'hémisphère a pour surface $2\pi R^2$: l'aire d'une sphère de rayon R est donc $4\pi R^2$.

Intégrale en coordonnées cylindriques. Si $M = (x, y, z)$ est un point de l'espace non situé sur l'axe Oz, les *coordonnées cylindriques* de M sont (r, θ, z), où r et θ sont les coordonnées polaires du point $m = (x, y)$ projection de M sur le plan xOy.
Si W est l'ensemble des coordonnées cylindriques d'un domaine V de l'espace, la formule du changement de variables s'écrit



$$\iiint_V f(x,y,z)\,dxdydz = \iiint_W f(r\cos\theta, r\sin\theta, z)\,r\,drd\theta dz$$

Remarque
Pour le changement de variables des coordonnées sphériques, le déterminant de la matrice jacobienne est $r^2\cos\varphi$ (page 368).

1.3 Intégrales doubles généralisées

On doit parfois considérer l'intégrale d'une fonction $f(x, y)$ sur un domaine non borné, comme par exemple un quart de plan $D = \{(x,y) \in \mathbb{R}^2 \mid x \geqslant a \text{ et } y \geqslant b\}$.
Posons $D_{u,v} = \{(x,y) \mid a \leqslant x \leqslant u, b \leqslant y \leqslant v\}$. Si les intégrales $I(u,v) = \iint_{D_{u,v}} f(x,y)\,dxdy$ ont une limite quand u et v tendent vers $+\infty$, cette limite est l'intégrale généralisée $\iint_D f(x,y)\,dxdy$.
Voici des conditions qui assurent l'existence de l'intégrale généralisée.
Pour que $\iint_D f(x,y)\,dxdy$ *existe, il suffit que pour tout* $x \geqslant a$, *l'intégrale généralisée* $\varphi(x) = \int_b^{+\infty} |f(x,y)|\,dy$ *existe et que l'intégrale généralisée* $\int_a^{+\infty}\varphi(x)\,dx$ *existe.*

Soit f une fonction définie sur le quart de plan $D = \{(x,y) \mid x \geqslant 0 \text{ et } y \geqslant 0\}$. Passons en coordonnées polaires en posant $F(r,\theta) = f(r\cos\theta, r\sin\theta)$ et notons D_R le secteur formé des points de D tels que $r \leqslant R$. Si elle existe, l'intégrale généralisée $I = \iint_D f(x,y)\,dxdy$ est la limite quand R tend vers $+\infty$ des intégrales $I_R = \iint_{D_R} f(x,y)\,dxdy = \int_0^{\pi/2}\left(\int_0^R rF(r,\theta)dr\right)d\theta$.

Critère d'existence. *S'il existe un nombre* $\alpha > 2$ *et une constante* k *tels que* $\left|f(x,y)\right| \leqslant \frac{k}{r^\alpha}$*, alors* I *existe.*

En effet, on a alors $\left|rF(r,\theta)\right| \leqslant \frac{k}{r^{\alpha-1}}$ et $\alpha - 1 > 1$, donc $\int_0^{+\infty} rF(r,\theta)\, dr$ existe (page 325).

Exemples. Le domaine D est le quart de plan $x \geqslant 0,\ y \geqslant 0$.

► Si $\alpha > 1$, l'intégrale généralisée $\iint_D \frac{\cos y\, dxdy}{(x^2+y^2+1)^\alpha}$ existe.

En passant aux coordonnées polaires, on a en effet $\left|f(x,y)\right| = \frac{1}{(r^2+1)^\alpha} \leqslant \frac{1}{r^{2\alpha}}$ et l'on a $2\alpha > 2$ si $\alpha > 1$.

► On a $I = \iint_D e^{-(x^2+y^2)}\, dxdy = \frac{\pi}{4}$.

En effet, $I = \lim_{R\to+\infty} I_R$, où

$$I_R = \int_0^{\pi/2}\int_0^R e^{-r^2} r\, drd\theta = \frac{\pi}{2}\int_0^R re^{-r^2}\, dr = \frac{\pi}{2}\left[-\frac{1}{2}e^{-u}\right]_0^{R^2} = \frac{\pi}{4}(1 - e^{-R^2})$$

► Puisque $e^{-(x^2+y^2)} = e^{-x^2}e^{-y^2}$, on a aussi, en utilisant la propriété (d) page 399,

$$I = \iint_D e^{-(x^2+y^2)}\, dxdy = \left[\int_0^{+\infty} e^{-x^2}\, dx\right]\left[\int_0^{+\infty} e^{-y^2}\, dy\right] = J^2, \text{ avec } J = \int_0^{+\infty} e^{-x^2}\, dx.$$

Il vient donc $\int_0^{+\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}$. Par le changement de variable $x = t/\sqrt{2}$, on en déduit $\int_0^{+\infty} e^{-t^2/2}\, dt = \frac{1}{2}\sqrt{2\pi}$ et la valeur de l'intégrale de Gauss (page 333)

$$\int_{-\infty}^{+\infty} e^{-t^2/2}\, dt = \sqrt{2\pi}$$

2. Application aux probabilités

Variable aléatoire à valeurs dans $\mathbb{R}^2$

Soient X et Y des variables aléatoires définies sur le même ensemble d'événements. On note (X,Y) la variable aléatoire dont la valeur sur tout événement ω est le couple $(x,y) = \big(X(\omega), Y(\omega)\big)$.

La variable aléatoire (X,Y) *prend ses valeurs dans l'ensemble* $\mathbb{R}^2$ *des couples de nombres réels.*

La *fonction de répartition* de la variable aléatoire $V = (X,Y)$ est définie par

$$H(x,y) = P\big[(X \leqslant x) \text{ et } (Y \leqslant y)\big]\ , \text{ pour tout } (x,y) \in \mathbb{R}^2.$$

Il s'ensuit que les variables X et Y sont indépendantes si et seulement si la fonction de répartition de (X,Y) est le produit $F_X(x)F_Y(y)$, où F_X et F_Y sont les fonctions de répartition de X et Y (définition page 335).

Pour tout couple (x,y) de nombres réels, posons $D_{x,y} =]-\infty, x] \times]-\infty, y]$. S'il existe une fonction h de deux variables, continue et positive, telle que pour tout (x, y),

$$H(x,y) = \iint_{D_{x,y}} h(t,u)\, dtdu,$$

alors h s'appelle la *densité* de (X,Y). Dans ce cas, on a $\dfrac{\partial^2 H}{\partial x \partial y} = h$ et

$$P\big[(a < X \leqslant b) \text{ et } (c < Y \leqslant d)\big] = \iint_R h(t,u)\, dtdu, \text{ où } R \text{ est le rectangle }]a,b] \times]c,d].$$

Espérance et variance pour des variables indépendantes

Soient X et Y deux variables aléatoires possédant une espérance et une variance. Notons $V(X)$ la variance de X et rappelons que la covariance de (X,Y) est $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$ (page 381).

Proposition. *Si X et Y sont indépendantes, alors on a $E(XY) = E(X)E(Y)$, $V(X+Y) = V(X) + V(Y)$ et $\mathrm{Cov}(X,Y) = 0$.*

Justification. Montrons la formule pour l'espérance du produit XY dans le cas particulier où l'ensemble des événements est fini. Notons $x_1, x_2, \ldots, x_p$ et $y_1, y_2, \ldots, y_n$ les valeurs possibles de X et de Y et $\omega_{i,j}$ l'événement $(X = x_i$ et $Y = y_j)$. Par définition, l'espérance de XY est $E(XY) = \sum_{i,j} x_i y_j P(\omega_{i,j})$. On a $P(\omega_{i,j}) = P(X = x_i)P(Y = y_j)$ car X et Y sont indépendantes, d'où $E(XY) = \Big(\sum_i x_i P(X = x_i)\Big)\Big(\sum_j y_j P(Y = y_j)\Big) = E(X)E(Y)$. Par définition de la covariance, on en déduit $\mathrm{Cov}(X,Y) = 0$. La variance de $Z = X+Y$ est l'espérance de $Z^2 - \big[E(Z)\big]^2 = X^2 - \big[E(X)\big]^2 + Y^2 - \big[E(Y)\big]^2 + 2\big(XY - E(X)E(Y)\big)$. On a ainsi la formule générale $V(X+Y) = V(X) + V(Y) + 2\,\mathrm{Cov}(X,Y)$, d'où le résultat. ■

Estimation empirique de la variance

Soit X une variable aléatoire et $(X_1, X_2, \ldots, X_n)$ une suite de réalisations indépendantes suivant la loi de X. On sait que la moyenne $Y = \dfrac{1}{n}(X_1 + \cdots + X_n)$ est une estimation de l'espérance de X : en effet, l'espérance est une fonction linéaire (pages 70 et 329) et toutes les variables X_i ont pour espérance $E(X_1)$, donc $E(Y) = E(X_1)$. Pour estimer la variance $E(X^2) - \big[E(X)\big]^2$ de X, considérons donc

$$\bar{v} = \frac{1}{n}(X_1^2 + \cdots + X_n^2) - Y^2$$

Puisqu'on a $E(X_i^2) = E(X_1^2)$, il vient $E\left(\frac{1}{n}(X_1^2 + \cdots + X_n^2)\right) = \frac{1}{n}nE(X_1^2) = E(X_1^2)$.
D'autre part, $Y^2 = \frac{1}{n^2}\left(\sum_i X_i^2 + 2\sum_{i<j} X_iX_j\right)$, d'où

$$\begin{aligned}
E(Y^2) &= \frac{1}{n^2}nE(X_1^2) + \frac{2}{n^2}\sum_{i<j} E(X_i)E(X_j) \quad \text{(indépendance des } X_i\text{)} \\
&= \frac{1}{n}E(X_1^2) + \frac{2}{n^2}\,\frac{n(n-1)}{2}\big[E(X_1)\big]^2\ , \quad \text{car } E(X_i) = E(X_1) \\
&= \frac{1}{n}E(X_1^2) + \frac{n-1}{n}\big[E(X_1)\big]^2
\end{aligned}$$

Finalement, $E(\bar{v}) = E(X_1^2) - \frac{1}{n}E(X_1^2) - \frac{n-1}{n}\big[E(X_1)\big]^2 = \frac{n-1}{n}\left(E(X_1^2) - \big[E(X_1)\big]^2\right)$, donc $E(\bar{v}) = \frac{n-1}{n}v$, où v est la variance de X_1.

La variable $\frac{n}{n-1}\,\bar{v}$ *est une estimation empirique de la variance de* X.

Loi d'une somme

Soient X et Y deux variables aléatoires indépendantes de densité f et g. Selon la définition donnée page 335, la fonction de répartition de $X+Y$ est définie par la formule $H(s) = P\big(X + Y \leqslant s\big) = \iint_{x+y\leqslant s} f(x)g(y)\,dxdy$, ou encore

$$H(s) = \int_{-\infty}^{+\infty} f(x)\,P\big(Y \leqslant s-x\big)\,dx = \int_{-\infty}^{+\infty}\left(f(x)\int_{-\infty}^{s-x} g(y)\,dy\right)dx$$

Faisons le changement de variable $t = y + x$ dans la dernière intégrale : on a $\int_{-\infty}^{s-x} g(y)\,dy = \int_{-\infty}^{s} g(t-x)\,dt$, d'où $H(s) = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{s} f(x)g(t-x)\,dt\right)dx$ et en intervertissant l'ordre des intégrations, il vient

$$H(s) = \int_{-\infty}^{s}\left(\int_{-\infty}^{+\infty} f(x)g(t-x)\,dx\right)dt$$

Par définition, la fonction $h(t) = \int_{-\infty}^{+\infty} f(x)g(t-x)\,dx$ est donc la densité de $X+Y$.

Proposition. *Si X et Y sont deux variables aléatoires indépendantes ayant pour densité f et g, alors $X+Y$ a pour densité $h(t) = \int_{-\infty}^{+\infty} f(t)g(t-x)\,dx$.*

Lorsqu'on fait des mesures, les résultats sont souvent distribués autour de leur moyenne selon une loi normale. Des erreurs d'origine diverses pouvant s'ajouter, on a besoin de savoir comment se comporte leur somme. Si les erreurs sont indépendantes, nous avons vu que la variance de la somme est la somme des variances, mais voici un résultat plus précis.

Somme de deux variables normales indépendantes. *Soient X et Y des variables aléatoires indépendantes suivant les lois normales $\mathcal{N}(m_1,\sigma_1)$ et $\mathcal{N}(m_2,\sigma_2)$. Alors $X+Y$ suit la loi normale $\mathcal{N}(m,\sigma)$, où $m=m_1+m_2$ et $\sigma=\sqrt{\sigma_1^2+\sigma_2^2}$.*

Démonstration. En remplaçant X et Y par $X-m_1$ et $Y-m_2$, on se ramène au cas $m_1=m_2=0$. Par définition de la loi normale (page 334) et d'après la proposition, la densité de $X+Y$ est dans ce cas

$$h(t)=\frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{+\infty}\exp\left(-\frac{x^2}{2\sigma_1^2}\right)\exp\left(-\frac{(t-x)^2}{2\sigma_2^2}\right)dx$$

En développant le carré dans la seconde exponentielle, la fonction sous le signe intégrale s'écrit

$$\begin{aligned}\exp\left(-\frac{t^2}{2\sigma_2^2}\right)\exp\left(-\frac{x^2\sigma^2}{2\sigma_1^2\sigma_2^2}\right)\exp\left(\frac{tx}{\sigma_2^2}\right)&=\exp\left(-\frac{t^2}{2\sigma_2^2}\right)\exp\left(-\frac{(x\sigma-t\sigma_1^2/\sigma)^2}{2\sigma_1^2\sigma_2^2}\right)\exp\left(\frac{t^2\sigma_1^2}{2\sigma^2\sigma_2^2}\right)\\&=\exp\left(\frac{t^2(\sigma_1^2-\sigma^2)}{2\sigma^2\sigma_2^2}\right)\exp\left(-\frac{(x\sigma-t\sigma_1^2/\sigma)^2}{2\sigma_1^2\sigma_2^2}\right)=\exp\left(-\frac{t^2}{2\sigma^2}\right)\exp\left(-\frac{(x\sigma-t\sigma_1^2/\sigma)^2}{2\sigma_1^2\sigma_2^2}\right)\end{aligned}$$

On a donc $h(t)=\dfrac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\dfrac{t^2}{2\sigma^2}\right)\displaystyle\int_{-\infty}^{+\infty}\exp\left(-\dfrac{(x\sigma-t\sigma_1^2/\sigma)^2}{2\sigma_1^2\sigma_2^2}\right)dx$. Par le changement de variable $u=x\sigma$, cette dernière intégrale s'écrit $\dfrac{1}{\sigma}\displaystyle\int_{-\infty}^{+\infty}\exp\left(-\dfrac{(u-a)^2}{2\sigma_1^2\sigma_2^2}\right)du$, où a est une constante, donc est égale à $\dfrac{1}{\sigma}\sqrt{2\pi}\sigma_1\sigma_2$ par définition de la loi de répartition normale. On obtient ainsi l'égalité $h(t)=\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{t^2}{2\sigma^2}\right)$. ■

3. Produit de convolution

Définition

Soient f et g des fonctions. Si l'intégrale généralisée $h(t)=\int_{-\infty}^{+\infty}f(x)g(t-x)\,dx$ existe, la fonction h s'appelle le *produit de convolution* des fonctions f et g et se note $f*g$.

► Si le produit de convolution de f et g existe, alors $f*g=g*f$.

Par le changement de variable $u=t-x$, il vient en effet

$$(f*g)(t)=\int_{-\infty}^{+\infty}f(x)g(t-x)\,dx=\int_{+\infty}^{-\infty}-f(t-u)g(u)\,du=\int_{-\infty}^{+\infty}g(u)f(t-u)\,du=(g*f)(t)$$

► Si a_1 et a_2 sont des nombres et si les produits de convolution f_1*g et f_2*g existent, alors $(a_1f_1+a_2f_2)*g=a_1(f_1*g)+a_2(f_2*g)$.

► Si l'une des fonctions f ou g vaut 0 en dehors d'un segment $[a,b]$, alors le produit de convolution $f*g$ existe.

En effet, si l'on a $f(x)=0$ pour x hors de $[a,b]$, l'intégrale généralisée est l'intégrale ordinaire $\int_a^b f(x)g(t-x)\,dx$.

- De nombreuses mesures physiques conduisent à l'analyse d'un *signal causal* : c'est une fonction u telle que $u(x)=0$ pour tout $x<0$. Si u et v sont des signaux causaux, leur produit de convolution $u*v$ existe et l'on a $(u*v)(t)=\int_0^t u(x)v(t-x)\,dx$.

  On a en effet $v(x)=0$ si $x<0$ et $u(t-x)=0$ si $x>t$, donc $u(x)v(t-x)=0$ si x est en dehors du segment $[0,t]$.

- Pour trois fonctions f,g,h, on a l'égalité $f*(g*h)=(f*g)*h$ si tous les produits de convolution existent.

Produit de convolution en théorie du signal

Un analyseur linéaire est un système électronique de traitement des signaux fonctionnant sur le mode

$$\text{signal d'entrée } u \quad \longmapsto \quad \text{réponse } u*f$$

La fonction f est caractéristique du dispositif. En entrée, on envoie le plus souvent un signal causal u dont la variable est le temps.

Réponse à une impulsion. Prenons comme signal d'entrée une fonction u_h constante sur le segment $[a,a+h]$ et nulle hors de cet intervalle ; pour que le signal d'entrée soit toujours de moyenne égale à 1, fixons $1/h$ comme valeur de u_h sur $[a,a+h]$. On a donc

$$(u_h*f)(t)=\int_{-\infty}^{+\infty}u_h(x)f(t-x)\,dx=\int_a^{a+h}u_h(x)f(t-x)\,dx=\frac{1}{h}\int_a^{a+h}f(t-x)\,dx$$

L'intégrale de droite est $\dfrac{\varphi(h)}{h}$, où $\varphi(h)=\int_a^{a+h}f(t-x)\,dx$. Si h tend vers 0, la réponse $\frac{1}{h}\int_a^{a+h}f(t-x)\,dx$ à l'instant t tend donc vers $\varphi'(0)=f(t-a)$: physiquement, cela signifie que si le signal d'entrée est une impulsion unité à l'instant a, la réponse est la fonction $f_a(t)=f(t-a)$, c'est-à-dire la fonction f affectée d'un retard a (à l'instant $t=a$, sa valeur est $f_a(a)=f(0)$).

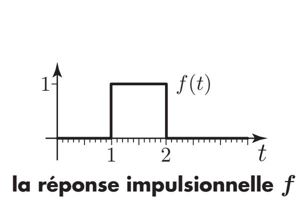
En particulier, si l'entrée est l'impulsion unité à l'instant 0, la réponse est f : on dit que f est *la réponse impulsionnelle* du système.

La réponse à une impulsion $v(x)$ à l'instant x est le signal $t\mapsto v(x)f_x(t)$. Puisque $(v*f)(t)=\int_{-\infty}^{+\infty}v(x)f_x(t)\,dx$, le produit de convolution $v*f$ apparaît comme « la somme sur x » des réponses impulsionnelles $t\mapsto f_x(t)$ pondérées par la valeur du signal à l'instant x.

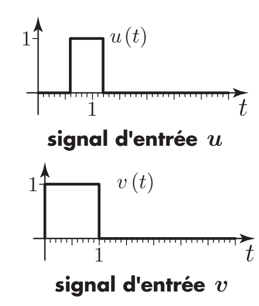
Exemple. Supposons que la réponse impulsionnelle est définie par $f(x)=1$ si $1<x<2$ et $f(x)=0$ sinon. Pour le signal d'entrée défini par $u(t)=1$ si $0{,}6<t<1{,}2$

et $u(t) = 0$ sinon, la réponse est

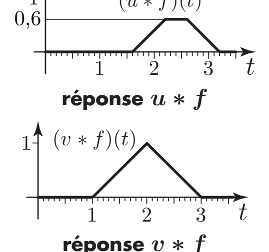
$$(u * f)(t) = \begin{cases} 0 & \text{si } t \leqslant 1{,}6 \qquad t-1{,}6 \quad \text{si } 1{,}6 < t \leqslant 2{,}2 \\ 0{,}6 & \text{si } 2{,}2 < t \leqslant 2{,}6 \qquad 3{,}2-t \quad \text{si } 2{,}6 < t \leqslant 3{,}2 \\ 0 & \text{si } t > 3{,}2 \end{cases}$$



la réponse impulsionnelle f



signal d'entrée v



réponse $v * f$

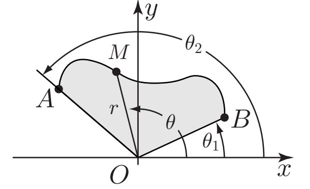
Ces signaux d'entrée présentent des discontinuités, mais les réponses sont continues sur $\mathbb{R}$: le produit de convolution a un effet régularisant.

D'après la proposition du paragraphe précédent, la fonction $u * f$, par exemple, est aussi la densité de probabilité de la somme $X + Y$ de deux variables aléatoires de densité uniforme sur les intervalles $[0{,}6\ ,\ 1{,}2]$ et $[1, 2]$.

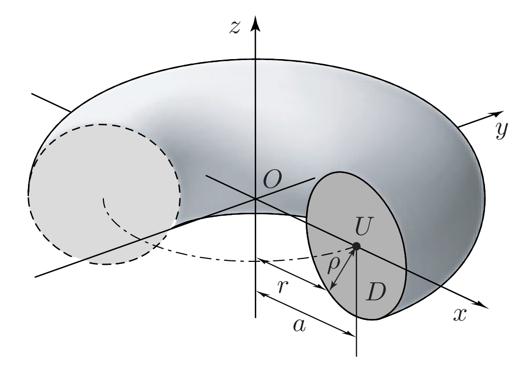
Exercices

1. Aire d'un secteur plan. On considère un secteur plan délimité par deux segments de droites OA et OB issus de l'origine et par une courbe joignant A et B d'équation $OM = r = f(\theta)$ en coordonnées polaires. Soit θ_1 l'angle $\widehat{Ox, \overrightarrow{OA}}$ et θ_2 l'angle $\widehat{Ox, \overrightarrow{OB}}$, où l'on suppose $0 \leqslant \theta_1 < \theta_2 \leqslant 2\pi$.

Montrer que l'aire du secteur est $\dfrac{1}{2}\displaystyle\int_{\theta_1}^{\theta_2} \big[f(\theta)\big]^2\, d\theta$.



2. Volume d'un tore plein. Faisons tourner un disque D de rayon ρ autour d'un axe situé dans le plan de D et situé à une distance $a > \rho$: le volume de l'espace ainsi obtenu s'appelle un tore plein (l'intérieur d'une chambre à air en est une bonne image). On veut calculer le volume V ainsi obtenu. Appelons U le centre du disque, Oz l'axe de révolution, l'origine O étant le projeté orthogonal de U sur l'axe ; l'axe Ox est dirigé par $\overrightarrow{OU}$ et Oy est orthogonal au plan du disque. Le point U a donc pour coordonnées cylindriques $(r = a, \theta = 0, z = 0)$.



a) Montrer que le disque est formé des points dont les coordonnées vérifient $\theta = 0$ et $z^2 + (a-r)^2 \leqslant \rho^2$. Notons Δ l'ensemble des couples (r, z) vérifiant cette inégalité.

b) Montrer que $V = 2\pi \iint_\Delta r\,dr\,dz$. En déduire $V = 2\pi \int_{-\rho}^{\rho} \left(\int_{u(z)}^{v(z)} r\,dr \right) dz$, où $u(z) = a - \sqrt{\rho^2 - z^2}$ et $v(z) = a + \sqrt{\rho^2 - z^2}$.

c) Montrer que $V = 2a\pi^2\rho^2$.

@ **3. Aire d'une surface torique.** Reprenons les notations de l'exercice précédent et appelons S la surface du tore.

a) Pour tout point M du plan xOz, notons u l'angle $\widehat{\overrightarrow{Ox}, \overrightarrow{OM}}$. Montrer que le bord du disque D est formé des points tels que $x = a + \rho \cos u$ et $z = \rho \sin u$. En déduire que S est formée des points $M(u, \theta)$ de coordonnées cartésiennes $x = (a + \rho \cos u) \cos \theta$, $y = (a + \rho \cos u) \sin \theta$, $z = \rho \sin u$, où $0 \leqslant u \leqslant 2\pi$ et $0 \leqslant \theta \leqslant 2\pi$.

b) Calculer les vecteurs $\dfrac{\partial M}{\partial u}$ et $\dfrac{\partial M}{\partial \theta}$ et montrer que l'élément d'aire est $\rho(a + \rho \cos u)\,du\,d\theta$. En déduire que l'aire du tore est $4\pi^2\rho a$.

@ **4.** Soient X et Y des variables aléatoires indépendantes et de loi uniforme sur $[0, 1]$. On pose $M = \max\{X, Y\}$ et $m = \min\{X, Y\}$. Soit t un nombre compris entre 0 et 1.

a) Montrer que l'on a $P(M < t) = t^2$ et que l'espérance de M est $2/3$. Montrer que $P(m < t) = t(2-t)$ et que l'espérance de m est $1/3$ (comparer à l'exercice **10** page 340).

b) Pour tous nombres u et v entre 0 et 1, notons $f(u,v)$ la probabilité de l'événement $\big[(M<u)$ et $(m<v)\big]$. Montrer que $f(u,v) = v(2u - v)$ et que la densité de (M,m) est 2.

c) On pose $D = M - m$. Montrer que $P(D < t) = 2 \iint_D du\,dv$, où $D = \{\,(u,v) \mid 0 \leqslant u \leqslant 1$ et $u - t \leqslant v \leqslant u\}$. En déduire que $P(D < t) = t(2-t)$. Quelle est l'espérance de D ?

d) On pose $Q = m/M$. Montrer que $P(Q < t) = 2 \iint_\Delta du\,dv$, où $\Delta = \{\,(u,v) \mid 0 \leqslant u \leqslant 1$ et $v \leqslant tu\,\}$. En déduire que $P(Q < t) = t$.

@ **5.** On fait un essai d'engrais sur onze parcelles de terrain : sur les cinq premières, on met l'engrais A et sur les six dernières, l'engrais B. Le tableau ci-contre donne les rendements observés.

engrais A					engrais B					
1	2	3	4	5	6	7	8	9	10	11
17	12	25	28	22	23	29	19	25	15	27

On veut savoir à quel risque on peut affirmer que B a un meilleur rendement que A. Pour cela, considérons les rendements respectifs comme des variables aléatoires R_A et R_B et faisons l'hypothèse que ces variables sont indépendantes et qu'ils suivent une loi normale. Notons M_A et M_B les rendements moyens de A et de B (ce sont aussi des variables aléatoires).

a) Calculer l'espérance empirique m_A et la variance empirique v_A de R_A. Calculer de même l'espérance empirique m_B et la variance empirique v_B de R_B. (On trouve $m_A = 20{,}8$, $v_A = (5/4) \times 32{,}56$, $m_B = 23$ et $v_B = (6/5) \times 22{,}66$.)

b) Montrer que la variance empirique de M_A est $\sigma_A^2 = \dfrac{1}{5^2} 5v_A = \dfrac{v_A}{5}$ et que la variance empirique de M_B est $\sigma_B^2 = \dfrac{v_B}{6}$. En déduire que l'écart-type empirique de $D = M_A - M_B$ est $\sigma = \sqrt{\dfrac{v_A}{5} + \dfrac{v_B}{6}}$. Quelle est l'espérance empirique de D ?

c) Montrer que la probabilité pour que $D < m_A - m_B + b\sigma$ vaut à peu près $\Phi(b)$, où Φ est la fonction de Gauss (page 335). Calculer b pour que $m_A - m_B + b\sigma = 0$. En déduire qu'on peut accepter l'hypothèse « l'engrais B a un meilleur rendement que l'engrais A » au risque 27% environ.

6. Produit de convolution et approximation en moyenne. Si f est une fonction continue dont la valeur absolue $|f(x)|$ a une intégrale généralisée de $-\infty$ à $+\infty$, on pose $\|f\| = \int_{-\infty}^{+\infty} |f(x)|\, dx$. Supposons que f et g sont de telles fonctions. Alors le produit de convolution $\big(|f| * |g|\big)(t) = \int_{-\infty}^{+\infty} |f(x)|\, |g(t-x)|\, dx$ existe quel que soit t, résultat que nous admettons.

a) Montrer que $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |f(x)|\, |g(t-x)|\, dx\, dt \leqslant \iint_{\mathbb{R}^2} |f(x)|\, |g(u)|\, dx\, du$ (faire le changement de variable $u = t-x$).

b) Remarquer que $|f * g(t)| \leqslant \int_{-\infty}^{+\infty} |f(x) g(t-x)|\, dx$. En déduire l'inégalité

$$\|f * g\| \leqslant \|f\|\, \|g\|\,.$$

c) On dit que des fonctions f_n *tendent en moyenne vers* f si $\|f - f_n\|$ tend vers 0 quand n tend vers l'infini. En utilisant **(b)**, montrer que si f_n tend en moyenne vers f, alors $f_n * g$ tend en moyenne vers $f * g$.

Chapitre 14

Champ de vecteurs, formes différentielles

1. Champ de vecteurs

Repérons les points dans le plan ou l'espace au moyen du repère orthonormé $(O;\vec{i},\vec{j},\vec{k})$. Un champ de vecteurs est la donnée, en tout point M d'une région de l'espace (ou du plan xOy), d'un vecteur $\overrightarrow{E(M)}$ de l'espace (ou du plan) dépendant du point M.

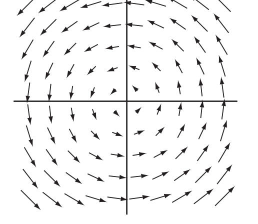
Définitions

- Si U est un domaine de l'espace, un *champ de vecteurs* sur U est une fonction $\overrightarrow{E(x,y,z)} = P(x,y,z)\,\vec{i} + Q(x,y,z)\,\vec{j} + R(x,y,z)\vec{k}$, où P, Q, R sont des fonctions à valeurs réelles définies dans U.
- Si D est un domaine plan, une fonction $\overrightarrow{E(x,y)} = P(x,y)\,\vec{i} + Q(x,y)\vec{j}$ est un champ de vecteurs sur D.

Nous supposons dans la suite que les fonctions coordonnées P, Q, R ont des dérivées secondes continues.

Exemples

Un champ de vitesses. Dans un fluide en mouvement, chaque particule possède, à un instant donné, un vecteur vitesse v qui dépend de sa position : si la particule est au point de coordonnées (x, y, z), sa vitesse est $v(x,y,z) = \dfrac{\partial x}{\partial t}\vec{i} + \dfrac{\partial y}{\partial t}\vec{j} + \dfrac{\partial z}{\partial t}\vec{k}$.
La figure ci-contre montre le champ des vitesses dans un mouvement circulaire plan autour de l'origine : en tout point $M = (x,y)$, le vecteur $v(x,y)$ est orthogonal à $\overrightarrow{OM}$, donc de la forme $v(x,y) = \omega(x,y)(-y,x)$. Le nombre $\omega(x,y)$ est la



vitesse angulaire car $\|v\| = |\omega|\sqrt{(-y)^2 + x^2} = r|\omega|$, où $r = OM$ est la distance au centre. Sur la figure, nous avons choisi ω constant.

Le champ de gravitation. Une masse m_0 en un point A de l'espace exerce sur toute autre masse m placée en un point M distant de r, une force d'attraction $g(M) = -\frac{Gm_0 m}{r^2}\vec{u}$, où $\vec{u}$ est le vecteur unitaire dans la direction $\overrightarrow{AM}$. Le nombre positif G est la constante d'attraction. La fonction g est un champ de vecteurs. Les vecteurs $g(M)$ étant toujours dirigés vers le point A, on dit que le champ est *central*.

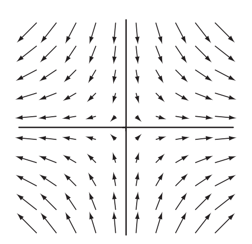
Le champ électrique. De même, une charge électrique q_0 en A exerce sur toute autre charge placée en un point M distant de r, une force $\overrightarrow{E(M)} = \frac{cq_0 q}{r^2}\vec{u}$, où c est une constante, positive ou négative, dépendant des unités choisies.

Lignes de champ

Définition
Si $\vec{E}$ est un champ de vecteurs dans un domaine D du plan, une *ligne de champ* est une courbe paramétrée régulière tracée dans D, tangente en tout point M au vecteur $\overrightarrow{E(M)}$ et parcourue dans le sens du champ.

En posant $M(t) = \big(x(t), y(t)\big)$ et $\vec{E} = P\vec{i} + Q\vec{j}$, cela signifie que pour tout t, les vecteurs $\big(x'(t), y'(t)\big)$ et $\big(P[x(t), y(t)], Q[x(t), y(t)]\big)$ sont colinéaires et de même sens.

Exemple. Le champ $\overrightarrow{E(x,y)} = x\vec{i} - y\vec{j}$ est représenté ci-contre : sur l'axe des abscisses ($y = 0$), le champ est horizontal, sur l'axe des ordonnées ($x = 0$), le champ est vertical et aux autres points $M = (x, y)$, la pente du champ est $\frac{-y}{x} = -t$, où t est la pente de $\overrightarrow{OM}$.



Une ligne de champ $\big(x(t), y(t)\big)$ doit vérifier

$$0 = \begin{vmatrix} x(t) & x'(t) \\ -y(t) & y'(t) \end{vmatrix} = x(t)y'(t) + x'(t)y(t) = \frac{d}{dt}\big(x(t)y(t)\big)$$

L'équation d'une ligne de champ est donc $xy = c$, où c est une constante.

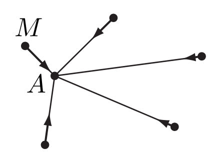
- Si $c = 0$, on obtient les quatre demi-axes de coordonnées.
- Pour $c \neq 0$, les lignes de champ sont des hyperboles ayant les axes pour asymptotes.

1.1 Champ de gradient

Soit $V(x, y, z)$ une fonction de trois variables, à valeurs réelles et ayant des dérivées continues.

Définition
Le champ de vecteurs $\overrightarrow{\text{Grad}}_V(x,y,z) = \dfrac{\partial V}{\partial x}\vec{i} + \dfrac{\partial V}{\partial y}\vec{j} + \dfrac{\partial V}{\partial z}\vec{k}$ s'appelle le *champ de gradient de V*. La fonction $\rho = -V$ est un *potentiel* du champ (il est défini à l'addition près d'une constante).

Exemple. Soit $\vec{E}$ un champ de vecteurs central dont l'intensité en tout point M est proportionnelle à $\dfrac{1}{r^2}$, r étant la distance de M au centre. En appelant A le centre, on a donc



$$\overrightarrow{E(M)} = \frac{c}{r^2}\vec{u}$$

où r est la distance AM, $\vec{u}$ le vecteur unitaire dans la direction $\overrightarrow{AM}$ et c une constante. Montrons que $\vec{E}$ est un champ de gradient, de potentiel $\rho(x,y,z) = \dfrac{c}{r}$.

Choisissons l'origine des coordonnées en A. Si $M = (x,y,z)$, alors $r^2 = x^2+y^2+z^2$, $\vec{u} = \frac{1}{r}\overrightarrow{AM} = \frac{1}{r}(x,y,z)$ et $\vec{E}(x,y,z) = \dfrac{cx}{r^3}\vec{i} + \dfrac{cy}{r^3}\vec{j} + \dfrac{cz}{r^3}\vec{k}$. On a $\dfrac{\partial r}{\partial x} = \dfrac{x}{r}$ et en posant $V = -\rho$, il vient $\dfrac{\partial V}{\partial x} = -c\left(-\dfrac{1}{r^2}\dfrac{\partial r}{\partial x}\right) = \dfrac{cx}{r^3}$; de même $\dfrac{\partial V}{\partial y} = \dfrac{cy}{r^3}$ et $\dfrac{\partial V}{\partial z} = \dfrac{cz}{r^3}$, donc $\vec{E}(x,y,z) = \overrightarrow{\text{Grad}}_V(x,y,z)$.

Propriétés du gradient

- $\overrightarrow{\text{Grad}}_{V+W} = \overrightarrow{\text{Grad}}_V + \overrightarrow{\text{Grad}}_W$.
- $\overrightarrow{\text{Grad}}_{VW}(M) = V(M)\,\overrightarrow{\text{Grad}}_W\,(M) + W(M)\,\overrightarrow{\text{Grad}}_V\,(M)$.

Expression du gradient en coordonnées cylindriques. Si (r,θ,z) sont les coordonnées cylindriques d'un point M, introduisons, comme page 368, les vecteurs orthogonaux unitaires $\vec{u} = \cos\theta\,\vec{i} + \sin\theta\vec{j}$ et $\vec{u_1} = -\sin\theta\,\vec{i} + \cos\theta\vec{j}$, de sorte qu'on a $\overrightarrow{OM} = r\,\vec{u} + z\vec{k}$. D'après le calcul fait au chapitre 12, l'expression du gradient de V est

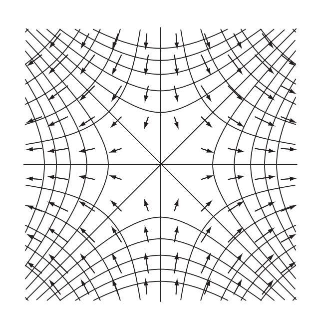
$$\overrightarrow{\text{Grad}}_V = \frac{\partial V}{\partial r}\,\vec{u} + \frac{1}{r}\frac{\partial V}{\partial \theta}\,\vec{u_1} + \frac{\partial V}{\partial z}\vec{k}$$

Lignes de champ d'un champ de gradient

Soit $\vec{E} = \overrightarrow{\text{Grad}}_V$ un champ de gradient dans le plan. En tout point M, la ligne de niveau de V passant par M est orthogonale au vecteur $\overrightarrow{\text{Grad}}_V(M)$ (page 370). Puisque le champ en M est $\overrightarrow{\text{Grad}}_V(M)$, cette ligne de niveau est orthogonale à la ligne de champ. Les lignes de niveau de V (ou de ρ, ce qui revient au même) s'appellent des *équipotentielles* et l'on a le résultat suivant.

Les lignes de champ de E sont les courbes orthogonales aux équipotentielles.

Exemple Reprenons le champ $E(x,y) = x\vec{i} - y\vec{j}$ (dernier exemple du précédent paragraphe). Si l'on pose $V(x,y) = \frac{1}{2}(x^2-y^2)$, alors $\frac{\partial V}{\partial x}=x$, $\frac{\partial V}{\partial y}=-y$, donc $\overrightarrow{\mathrm{Grad}}_V=(x,-y)=\vec{E}(x,y)$: ainsi le champ $\vec{E}$ est un champ de gradient, de potentiel $-V$. Les équipotentielles ont pour équation $x^2-y^2=k$, où k est une constante.



- Pour $k=0$, on obtient comme équipotentielles les droites d'équation $y=x$ et $y=-x$, bissectrices des axes.
- Si $k\neq 0$, l'équipotentielle d'équation $x^2-y^2=k$ est une hyperbole ayant pour asymptotes les bissectrices des axes.

En tout point $M=(a,b)$ du plan, il passe l'équipotentielle H d'équation $x^2-y^2=a^2-b^2$ et la ligne de champ L d'équation $xy=ab$. D'après le résultat précédent, H et L sont orthogonales au point M.

1.2 Rotationnel

Supposons que le champ de vecteur $\vec{E}(x,y,z)=P(x,y,z)\vec{i}+Q(x,y,z)\vec{j}+R(x,y,z)\vec{k}$ est un champ de gradient, de potentiel $-V$. Alors on a $(P,Q,R)=\left(\frac{\partial V}{\partial x},\frac{\partial V}{\partial y},\frac{\partial V}{\partial z}\right)$ et

$$\frac{\partial R}{\partial y}-\frac{\partial Q}{\partial z}=\frac{\partial}{\partial y}\left(\frac{\partial V}{\partial z}\right)-\frac{\partial}{\partial z}\left(\frac{\partial V}{\partial y}\right)=\frac{\partial^2 V}{\partial y\partial z}-\frac{\partial^2 V}{\partial z\partial y}=0$$

car on peut dériver dans l'ordre qu'on veut. De même,

$$\frac{\partial P}{\partial z}-\frac{\partial R}{\partial x}=\frac{\partial}{\partial z}\left(\frac{\partial V}{\partial x}\right)-\frac{\partial}{\partial x}\left(\frac{\partial V}{\partial z}\right)=0 \text{ et } \frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}=\frac{\partial}{\partial x}\left(\frac{\partial V}{\partial y}\right)-\frac{\partial}{\partial y}\left(\frac{\partial V}{\partial x}\right)=0.$$

Définition
Pour tout champ de vecteurs $\vec{E}=P\vec{i}+Q\vec{j}+R\vec{k}$, le champ de vecteurs

$$\overrightarrow{\mathrm{Rot}}(\vec{E})=\left(\frac{\partial R}{\partial y}-\frac{\partial Q}{\partial z}\right)\vec{i}+\left(\frac{\partial P}{\partial z}-\frac{\partial R}{\partial x}\right)\vec{j}+\left(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}\right)\vec{k}$$

s'appelle le *rotationnel de* $\vec{E}$.

énonçons ce que montre le calcul qu'on vient de faire.

Tout champ de gradient a un rotationnel nul.

Procédé de calcul du rotationnel. Définissons le « vecteur-opération »

$$\nabla=\left[\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right]\quad,\qquad \text{appelé } \textit{nabla}.$$

Pour tout champ de vecteur $\vec{E}$ de composantes $[P, Q, R]$, on obtient les composantes du rotationnel de E en calculant le « produit vectoriel » (page 223)

$$\nabla \wedge E = \begin{bmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{bmatrix} \wedge \begin{bmatrix} P \\ Q \\ R \end{bmatrix} = \begin{bmatrix} \dfrac{\partial R}{\partial y} - \dfrac{\partial Q}{\partial z} \\ \dfrac{\partial P}{\partial z} - \dfrac{\partial R}{\partial x} \\ \dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y} \end{bmatrix}$$

Pour un champ de vecteurs $\vec{E} = P(x,y)\,\vec{i} + Q(x,y)\vec{j}$ dans le plan, on a simplement

$$\overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\vec{k}$$

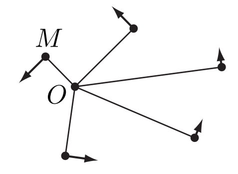
car la fonction R est nulle ainsi que les dérivées de P et de Q par rapport à z.

Le rotationnel d'un champ de vecteur dans le plan xOy est orthogonal à ce plan.

Exemple 1. Dans un mouvement plan circulaire uniforme de centre l'origine, le champ des vitesses $\vec{E}(x,y) = -\omega y\,\vec{i} + \omega x\vec{j}$ a un rotationnel constant :

$$\overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) = \left(\frac{\partial(\omega x)}{\partial x} - \frac{\partial(-\omega y)}{\partial y}\right)\vec{k} = 2\omega\vec{k}$$

Exemple 2. Soit D le domaine du plan formé des points différents de l'origine. Le champ de vecteurs sur D défini par $\overrightarrow{E(M)} = \frac{-y}{r^2}\vec{i} + \frac{x}{r^2}\vec{j}$, où r est la distance OM, a un rotationnel nul.



En effet, en posant $P = \dfrac{-y}{r^2}$ et $Q = \dfrac{x}{r^2}$, on a

$$\frac{\partial Q}{\partial x} = \frac{1}{r^2} + x\frac{\partial(r^{-2})}{\partial x} = \frac{1}{r^2} + x(-2r^{-3})\frac{\partial r}{\partial x} = \frac{1}{r^2} - \frac{2x}{r^3}\frac{x}{r} = \frac{1}{r^2} - \frac{2x^2}{r^4}$$

et de même $\dfrac{\partial P}{\partial y} = -\dfrac{1}{r^2} + \dfrac{2y^2}{r^4}$. Il vient

$$\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \frac{2}{r^2} - \frac{2x^2 + 2y^2}{r^4} = \frac{2}{r^2} - \frac{2r^2}{r^4} = 0\,.$$

Cependant, nous montrerons page 424 que $\vec{E}$ n'est pas un champ de gradient : avoir un rotationnel nul ne suffit pas toujours pour être un champ de gradient.

Formulaire

- $\overrightarrow{\mathrm{Rot}}(\,\overrightarrow{\mathrm{Grad}}_V\,) = \vec{0}$
- $\overrightarrow{\mathrm{Rot}}(\,\vec{E} + \vec{F}\,) = \overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) + \overrightarrow{\mathrm{Rot}}(\,\vec{F}\,)$
- Si f est une fonction, $\overrightarrow{\mathrm{Rot}}(f\,\vec{E}\,) = \overrightarrow{\mathrm{Grad}}_f \wedge \vec{E} + f\;\overrightarrow{\mathrm{Rot}}\,(\,\vec{E}\,)$.

► En coordonnées cylindriques, posons $\vec{E} = E_r\,\vec{u} + E_\theta\,\vec{u_1} + E_z\vec{k}$, où les vecteurs $\vec{u}$ et $\vec{u_1}$ sont définis page 368. L'expression du rotationnel est

$$\overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) = \left(\frac{1}{r}\frac{\partial E_z}{\partial\theta} - \frac{\partial E_\theta}{\partial z}\right)\vec{u} + \left(\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r}\right)\vec{u_1} + \left(\frac{\partial E_\theta}{\partial r} + \frac{E_\theta}{r} - \frac{1}{r}\frac{\partial E_r}{\partial\theta}\right)\vec{k}$$

Recherche d'un potentiel

Exemple. Considérons le champ $\vec{E} = \frac{-2x}{y}\vec{i} + \frac{x^2y+1}{y^3}\vec{j}$ défini sur le demi-plan D formé des points (x,y) tels que $y>0$. Les composantes du champ sont $P(x,y) = \frac{-2x}{y}$ et $Q(x,y) = \frac{x^2}{y^2} + \frac{1}{y^3}$, de sorte que $\frac{\partial P}{\partial y} = \frac{2x}{y^2} = \frac{\partial Q}{\partial x}$ et donc $\overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) = \vec{0}$.
Cherchons un potentiel $-V$, c'est-à-dire une fonction telle que

$$\frac{\partial V}{\partial x} = \frac{-2x}{y} \quad \text{et} \quad \frac{\partial V}{\partial y} = \frac{x^2}{y^2} + \frac{1}{y^3}$$

En intégrant en x, la première condition donne

$$V(x,y) = \frac{-x^2}{y} + u(y)\,,$$

où u est fonction de la seule variable y (voir page 369). Dérivons par rapport à y :

$$\frac{x^2}{y^2} + u'(y) = \frac{\partial V}{\partial y} = \frac{x^2}{y^2} + \frac{1}{y^3}$$

donc $u'(y) = \frac{1}{y^3}$ et $u(y) = -\frac{1}{2y^2} + k$, où k est une constante. Ainsi la fonction $V(x,y) = -\frac{x^2}{y} - \frac{1}{2y^2}$ a pour dérivées partielles $\frac{\partial V}{\partial x} = P$ et $\frac{\partial V}{\partial y} = Q$, autrement dit $\vec{E} = \overrightarrow{\mathrm{Grad}}_V$.

Proposition. *Soit $\vec{E}$ un champ de vecteurs défini dans un domaine D et tel que $\overrightarrow{\mathrm{Rot}}(\,\vec{E}\,) = \vec{0}$.*

a) *Supposons que pour tous points A et B de D, le parallélépipède (ou le rectangle) de diagonale AB à côtés parallèles aux axes est inclus dans D. Si (a,b,c) est un point de D, alors on a $\vec{E} = \overrightarrow{\mathrm{Grad}}_V$, où $V(x,y,z) = \int_a^x P(t,y,z)\,dt + \int_b^y Q(a,t,z)\,dt + \int_c^z R(a,b,t)\,dt$.*

b) *Supposons que pour tout point $M \in D$, le segment OM qui joint M à l'origine est inclus dans D. Alors on a $\vec{E} = \overrightarrow{\mathrm{Grad}}_V$, où*

$$V(x,y,z) = x\int_0^1 P(tx,ty,tz)\,dt + y\int_0^1 Q(tx,ty,tz)\,dt + z\int_0^1 Q(tx,ty,tz)\,dt.$$

L'espace tout entier, ou un demi-plan de frontière parallèle à un axe de coordonnée, possède la propriété (a) ; un disque ou un demi-plan contenant l'origine possède la propriété (b) ; l'espace ou le plan privé d'un point ne possède aucune de ces propriétés.

Démonstration. Prenons le cas d'un domaine plan et appelons P,Q les composantes de $\vec{E}$. Supposons la première condition vérifiée et raisonnons comme dans l'exemple ci-dessus : on

veut $\dfrac{\partial V}{\partial x} = P$, donc $V(x,y) = \int_a^x P(t,y)\,dt + u(y)$, où $u(y) = V(a,y)$.

$$\begin{aligned}\frac{\partial V}{\partial y}(x,y) &= \int_a^x \frac{\partial P}{\partial y}(t,y)\,dt + u'(y) \qquad \text{(dérivation sous le signe intégrale)}\\ &= \int_a^x \frac{\partial Q}{\partial x}(t,y)\,dt + u'(y) \qquad \text{car } \frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}\\ &= Q(x,y) - Q(a,y) + u'(y)\end{aligned}$$

Puisqu'on doit avoir $\dfrac{\partial V}{\partial y}(x,y) = Q(x,y)$, il s'ensuit $u'(y) = Q(a,y)$ et $u(y) = \int^y Q(a,t)\,dt$, ce qu'il fallait démontrer.

Supposons satisfaite la condition de (b) et posons $V(x,y) = x\int_0^1 P(tx,ty)\,dt + y\int_0^1 Q(tx,ty)\,dt$. Ces intégrales ont un sens, car si (x,y) est un point de D, alors pour tout nombre t entre 0 et 1, (tx,ty) est aussi dans D, par hypothèse. En dérivant sous le signe intégrale, on a

$$\begin{aligned}\frac{\partial V}{\partial x}(x,y) &= \int_0^1 P(tx,ty)\,dt + x\int_0^1 t\frac{\partial P}{\partial x}(tx,ty)\,dt + y\int_0^1 t\frac{\partial Q}{\partial x}(tx,ty)\,dt\\ &= \int_0^1 P(tx,ty)\,dt + x\int_0^1 t\frac{\partial P}{\partial x}(tx,ty)\,dt + y\int_0^1 t\frac{\partial P}{\partial y}(tx,ty)\,dt \qquad \text{car } \frac{\partial Q}{\partial x} = \frac{\partial P}{\partial y}\\ &= \int_0^1 P(tx,ty)\,dt + \int_0^1 tU(t)\,dt\end{aligned}$$

où $U(t) = x\dfrac{\partial P}{\partial x}(tx,ty) + y\dfrac{\partial P}{\partial y}(tx,ty) = \dfrac{d}{dt}P(tx,ty)$. Il vient en intégrant par parties $\int_0^1 tU(t)\,dt = \big[tP(tx,ty)\big]_0^1 - \int_0^1 P(tx,ty)\,dt = P(x,y) - \int_0^1 P(tx,ty)\,dt$, et donc $\dfrac{\partial V}{\partial x}(x,y) = P(x,y)$. On montre de même l'égalité $\dfrac{\partial V}{\partial y}(x,y) = Q(x,y)$. ■

Divergence d'un champ de vecteurs

Définition

La *divergence* du champ de vecteurs $\vec{E} = P\,\vec{i} + Q\,\vec{j} + R\vec{k}$ est la fonction $\operatorname{div}\vec{E} = \dfrac{\partial P}{\partial x} + \dfrac{\partial Q}{\partial y} + \dfrac{\partial R}{\partial z}$.

Formulaire

- $\operatorname{div}(\vec{E}+\vec{F}) = \operatorname{div}\vec{E} + \operatorname{div}\vec{F}$
- $\operatorname{div}\big[\overrightarrow{\mathrm{Rot}}(\vec{E})\big] = 0$
- $\operatorname{div}\big(\overrightarrow{\mathrm{Grad}}_V\big)$ est le *laplacien* $\Delta V = \dfrac{\partial^2 P}{\partial x^2} + \dfrac{\partial^2 Q}{\partial y^2} + \dfrac{\partial^2 R}{\partial z^2}$ de V.
- Si f est une fonction, $\operatorname{div}(f\vec{E}) = (f)\operatorname{div}\vec{E} + \overrightarrow{\mathrm{Grad}}_f \cdot \vec{E}$
- $\operatorname{div}(\vec{E}\wedge\vec{F}) = \big[\overrightarrow{\mathrm{Rot}}(\vec{E})\big]\cdot\vec{F} - \vec{E}\cdot\big[\overrightarrow{\mathrm{Rot}}(\vec{F})\big]$
- En coordonnées cylindriques, avec la notation $\vec{E} = E_r\,\vec{u} + E_\theta\,\vec{u_1} + E_z\vec{k}$, on a

$$\operatorname{div}\vec{E} = \frac{E_r}{r} + \frac{\partial E_r}{\partial r} + \frac{1}{r}\frac{\partial E_\theta}{\partial \theta} + \frac{\partial E_z}{\partial z}$$

Une application en hydrodynamique

Dans un fluide en mouvement, chaque particule a des coordonnées d'espace (x, y, z) et un vecteur vitesse $u\vec{i} + v\vec{j} + w\vec{k}$. Notons μ la masse volumique (qui dans le cas d'un gaz, par exemple, n'est pas nécessairement constante).

Considérons au point (x, y, z), un petit élément de volume $\delta x \delta y \delta z$ à côtés parallèles aux axes de coordonnées.

► Pendant un petit intervalle de temps δt, la masse de matière entrant dans ce volume par une face parallèle à yOz est

$$\mu(x, y, z) u(x, y, z)\, \delta x \delta y \delta z\, \delta t\,,$$

► et la quantité sortant par la face opposée est (au premier ordre)

$$\mu(x+\delta x, y, z) u(x+\delta x, y, z)\, \delta x \delta y \delta z\, \delta t = \left[(\mu u)(x, y, z) + \frac{\partial(\mu u)}{\partial x}(x, y, z)\, \delta x\right] \delta y \delta z\, \delta t$$

En raisonnant de même pour les autres faces, le bilan de masse est

$$\delta m = -\left(\frac{\partial(\mu u)}{\partial x} + \frac{\partial(\mu v)}{\partial y} + \frac{\partial(\mu w)}{\partial z}\right) \delta x \delta y \delta z\, \delta t$$

La masse $\mu\, \delta x \delta y \delta z$ de l'élément de volume varie de δm pendant le temps δt, donc

$$\frac{\delta m}{\delta t} = -\left(\frac{\partial(\mu u)}{\partial x} + \frac{\partial(\mu v)}{\partial y} + \frac{\partial(\mu w)}{\partial z}\right) \delta x \delta y \delta z = \frac{\delta \mu}{\delta t} \delta x \delta y \delta z$$

Puisque cette masse ne varie pas, en passant aux dérivées, on obtient

$$\frac{\partial(\mu u)}{\partial x} + \frac{\partial(\mu v)}{\partial y} + \frac{\partial(\mu w)}{\partial z} + \frac{\partial \mu}{\partial t} = 0 \quad \textit{(équation de continuité)}$$

Dans le cas d'une masse gazeuse thermiquement isolée, le paramètre naturel de contrôle est la pression p : si p_0 et μ_0 sont les valeurs de la pression et de la masse volumique à un certain instant, alors d'après la loi thermodynamique des gaz, on a $p/p_0 = (\mu/\mu_0)^\gamma$, où γ est une constante dépendant de la nature du gaz (pour l'air atmosphérique, γ vaut environ 1,4).

Supposons le fluide incompressible. Alors μ est constant et l'équation de continuité s'écrit $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$. En introduisant le champ des vitesses $\vec{\mathscr{V}} = u\vec{i} + v\vec{j} + w\vec{k}$, il vient

$$\operatorname{div} \vec{\mathscr{V}} = 0$$

Supposons enfin que les mouvements dans le fluide ne présentent aucun tourbillon : il n'y a aucune ligne de courant fermée, ce qui se traduit par $\overrightarrow{\mathrm{Rot}}(\mathscr{V}) = \vec{0}$. Si l'espace occupé est un domaine convenable (voir page 420), on sait que le champ des vitesses possède un potentiel ρ : on a $\vec{\mathscr{V}} = -\overrightarrow{\mathrm{Grad}}_\rho$. D'après l'équation de continuité, le laplacien de ρ est $\Delta\rho = \operatorname{div}(\overrightarrow{\mathrm{Grad}}_\rho) = -\operatorname{div} \vec{\mathscr{V}} = 0$, d'où

$$\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} + \frac{\partial^2 \rho}{\partial z^2} = 0\,.$$

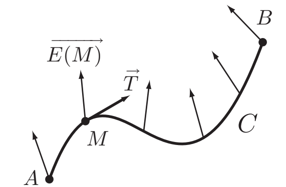
Dans ces conditions, le champ des vitesses dérive donc d'un potentiel harmonique. Pour avoir l'équation des lignes de courant, il faut trouver une fonction ρ harmonique

dans le domaine D et satisfaisant des conditions au bord de D ou à l'infini, comme par exemple $\overrightarrow{\text{Grad}}_\rho$ tangent au bord et ρ nul à l'infini (voir page 372 et l'exercice 2). Nous verrons page 433 que le mouvement du fluide est alors parfaitement déterminé.

1.3 Intégrale curviligne

Soit $\vec{E} = P\vec{i} + Q\vec{j} + R\vec{k}$ un champ de vecteurs et soit $C : M(t) = \big(x(t), y(t), z(t)\big)$ une courbe paramétrée régulière tracée dans le domaine du champ.

Interprétons le vecteur $\overrightarrow{E(M)}$ comme une force appliquée au point M. Le travail de la force le long d'un petit déplacement δs sur C, du point $M(t)$ au point $M(t+\delta t)$, est le produit scalaire $\overrightarrow{E(M)} \cdot \delta s\vec{T}$, où $\vec{T}$ est le vecteur tangent unitaire à C en M (seule compte la composante de la force dans la direction du déplacement). En paramétrant la courbe par l'abscisse curviligne s (page 313), le travail de $\vec{E}$ le long de C est donc



$$W = \int_0^L \overrightarrow{E\big(M(s)\big)} \cdot \overrightarrow{T(s)}\, ds, \quad \text{où } L \text{ est la longueur de } C.$$

On sait que $ds = \left\| \dfrac{d\overrightarrow{OM}}{dt} \right\| dt$, donc $\vec{T}ds = \dfrac{d\overrightarrow{OM}}{dt}\, dt$. Si le paramètre t varie de a à b, il vient $W = \int_a^b \overrightarrow{E\big(M(t)\big)} \cdot \dfrac{d\overrightarrow{OM}}{dt}\, dt$, ou encore

$$(*) \quad W = \int_a^b \Big[P\big(x(t),y(t),z(t)\big)x'(t) + Q\big(x(t),y(t),z(t)\big)y'(t) + R\big(x(t),y(t),z(t)\big)z'(t) \Big] dt$$

Définition

L'intégrale

$$\int_a^b \Big[P\big(x(t), y(t), z(t)\big)x'(t) + Q\big(x(t), y(t), z(t)\big)y'(t) + R\big(x(t), y(t), z(t)\big)z'(t) \Big] dt$$

s'appelle *la circulation* ou *l'intégrale du champ le long de* C et se note $\int_C P dx + Q dy + R dz$. On dit que c'est une *intégrale curviligne*.

Nous venons de montrer que la circulation du champ le long de C ne dépend pas de la façon dont on paramètre la courbe, à condition de garder le même sens de parcours. Mais si l'on change le sens de parcours, l'intégrale curviligne change de signe.

Calcul d'une intégrale curviligne. Pour calculer $W = \int_C P dx + Q dy + R dz$,

- on remplace x, y, z par $x(t)$, $y(t)$, $z(t)$ dans les fonctions P, Q, R,
- on remplace dx, dy, dz par les différentielles $x'(t)dt$, $y'(t)dt$, $z'(t)dt$,
- et l'intégrale curviligne est donnée par l'expression $(*)$.

Intégrale curviligne et potentiel

Supposons $\vec{E} = \overrightarrow{\mathrm{Grad}}_V$, donc $\left(\frac{\partial V}{\partial x}, \frac{\partial V}{\partial y}, \frac{\partial V}{\partial z}\right) = (P, Q, R)$. On a alors

$$dV = \frac{\partial V}{\partial x}\frac{dx}{dt}dt + \frac{\partial V}{\partial y}\frac{dy}{dt}dt + \frac{\partial V}{\partial z}\frac{dz}{dt}dt = Pdx + Qdy + Rdz$$

donc

$$\begin{aligned}\int_C Pdx+Qdy+Rdz &= \int_a^b \frac{d}{dt}V\big(x(t), y(t), z(t)\big)\,dt \\ &= V\big(x(b), y(b), z(b)\big) - V\big(x(a), y(a), z(a)\big) = V(B) - V(A)\end{aligned}$$

où A est l'origine de la courbe et B son extrémité.

Théorème. *Si $\vec{E}=\overrightarrow{\mathrm{Grad}}_V$, alors $\int_C Pdx+Qdy+Rdz=V(B)-V(A)$ pour toute courbe d'origine A et d'extrémité B.*

Si un champ possède un potentiel ρ, sa circulation le long d'une courbe est la différence de potentiel $\rho(A) - \rho(B)$ entre l'origine A et l'extrémité B : l'intégrale curviligne ne dépend pas de la forme de la courbe, mais seulement de la position des extrémités. Cette propriété est caractéristique d'un champ de gradient.

Conséquences. *Si un champ possède un potentiel,*

i) *son intégrale le long de toute courbe fermée est nulle ;*

ii) *aucune ligne de champ n'est fermée.*

Dans la direction du vecteur $\overrightarrow{\mathrm{Grad}}_V$, la fonction V est strictement croissante (page 365), donc le potentiel est décroissant. Puisqu'une ligne du champ $\vec{E} = \overrightarrow{\mathrm{Grad}}_V$ est en tout point tangente au gradient de V, le potentiel décroît strictement quand on parcourt cette ligne dans le sens du champ : si A est l'origine d'une ligne de champ et B l'extrémité, alors $\rho(A) - \rho(B)$ est un nombre strictement positif, donc non nul ; il s'ensuit que les points A et B sont différents, et la ligne de champ n'est pas fermée.

Exemple. Reprenons le champ $\vec{E}$ de composantes $P = \frac{-y}{r^2}$, $Q = \frac{x}{r^2}$, étudié dans l'exemple 2 page 419. Calculons la circulation de $\vec{E}$ le long du cercle C de rayon a centré à l'origine.
En paramétrant C par $x(t)=a\cos t, y=a\sin t$, il vient $Pdx+Qdy= \frac{-y\,dx + x\,dy}{r^2} = \frac{1}{a^2}(a^2\sin^2 t+a^2\cos^2 t)dt=dt$, d'où $\int_C Pdx+Qdy=2\pi$.
La circulation de $\vec{E}$ le long de la courbe fermée C n'est pas nulle, donc $\vec{E}$ n'est pas un champ de gradient.
Bien que son rotationnel soit nul, ce champ est défini dans le plan privé de l'origine, un domaine qui n'a aucune des propriétés énoncées dans la proposition page 420.

Exemple du champ de gravitation. Pour un champ de gravitation $\overrightarrow{E(M)} = -\frac{K}{r^2}\vec{u}$, où r est la distance OM et $\vec{u} = \frac{1}{r}\overrightarrow{OM}$, le potentiel est $\rho(M) = \frac{K}{r}$ (exemple page 417).
Pour toute courbe d'origine A et d'extrémité B, le travail du champ le long de cette courbe est $K\left(\frac{1}{OA} - \frac{1}{OB}\right)$.

Dans le cas de la gravité terrestre, pour de faibles variations d'altitude, on considère que le champ est constant : $\overrightarrow{E(M)} = -mg\vec{u}$, où g est la « constante » de gravitation terrestre et m la masse au point M ; on a alors $\overrightarrow{E(x,y,z)} = -mg\left(\frac{x}{r}, \frac{y}{r}, \frac{z}{r}\right) = -mg\overrightarrow{\mathrm{Grad}}_r = -\overrightarrow{\mathrm{Grad}}_{mgr}$ et le potentiel est mgr : le travail du champ de pesanteur le long d'une courbe est alors simplement le produit du poids mg par la différence d'altitude des extrémités.

La loi de conservation de l'énergie. Soit un point matériel M de masse m se déplaçant sous l'action d'un champ de force $\vec{E}$. Le vecteur vitesse est $\vec{v} = \frac{d\overrightarrow{OM}}{dt}$ et l'accélération est $\vec{\Gamma} = \frac{d\vec{v}}{dt}$. D'après la loi de Newton, on sait que $\vec{E} = m\vec{\Gamma}$. Si a et b sont les instants de départ et d'arrivée, le travail de la force le long de L est

$$\begin{aligned}
W &= \int_a^b \overrightarrow{E\big(M(t)\big)} \cdot \overrightarrow{v(t)}\, dt = \int_a^b m\, \overrightarrow{\Gamma(t)} \cdot \overrightarrow{v(t)}\ dt \\
&= m\int_a^b \big[\ddot{x}(t)\dot{x}(t) + \ddot{y}(t)\dot{y}(t) + \ddot{z}(t)\dot{z}(t)\big]\, dt \\
&= \frac{m}{2}\int_a^b \frac{d}{dt}\big[\dot{x}^2(t) + \dot{y}^2(t) + \dot{z}^2(t)\big]\, dt = \frac{m}{2}\int_a^b \frac{d}{dt}\Big(\|\ \overrightarrow{v(t)}\ \|^2\Big)\ dt \\
&= \frac{1}{2}m\|\ \overrightarrow{v(b)}\ \|^2 - \frac{1}{2}m\|\ \overrightarrow{v(a)}\ \|^2
\end{aligned}$$

La quantité $q(M) = \frac{1}{2}m\|\ \vec{v}\ \|^2$ est l'énergie cinétique : on a donc $W = q(B) - q(A)$, où A est le point de départ du mouvement et B le point d'arrivée.
Supposons que le champ de force possède un potentiel ρ. On a alors $W = \rho(A) - \rho(B)$, donc $q(B) - q(A) = \rho(A) - \rho(B)$ ou encore $q(A) + \rho(A) = q(B) + \rho(B)$.

Dans un mouvement sous l'action d'un champ de force ayant un potentiel, la somme de l'énergie cinétique et de l'énergie potentielle est constante.

2. Formule de Stokes

2.1 Formes différentielles

Définitions

Si P, Q, R sont des fonctions définies sur un domaine D de l'espace, l'expression $\omega = Pdx + Qdy + Rdz$ s'appelle une *forme différentielle de degré* 1 sur D. Le champ de vecteur $P\vec{i} + Q\vec{j} + R\vec{k}$ est le *champ associé à* ω.

Si $V(x,y,z)$ est une fonction, sa différentielle $dV = \frac{\partial V}{\partial x}dx + \frac{\partial V}{\partial y}dy + \frac{\partial V}{\partial z}dz$ est une forme différentielle de degré 1 et le champ associé à dV est $\overrightarrow{\text{Grad}}_V$.

Différentielle d'une forme différentielle

Dans la forme différentielle Pdx, remplaçons formellement P par sa différentielle et notons $d(Pdx)$ l'expression obtenue :

$$d(Pdx) = \left(\frac{\partial P}{\partial x}dx + \frac{\partial P}{\partial y}dy + \frac{\partial P}{\partial z}dz\right)dx = \frac{\partial P}{\partial x}dx\,dx + \frac{\partial P}{\partial y}dy\,dx + \frac{\partial P}{\partial z}dz\,dx$$

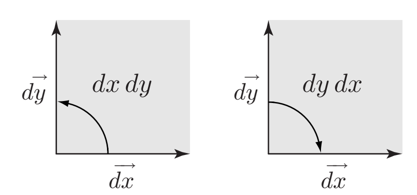
Sur les produits comme $dx\,dy$, on calcule comme s'il s'agissait d'un produit vectoriel de deux vecteurs :

► changer l'ordre des termes se traduit par un changement de signe :

$$dy\,dx = -dx\,dy \quad , \quad dz\,dx = -dx\,dz \quad , \quad dz\,dy = -dy\,dz$$

► on a les relations : $dx\,dx = dy\,dy = dz\,dz = 0$.

Justification : Ces règles s'expliquent si l'on pense que $dxdy$ est l'aire orientée d'un parallélépipède infinitésimal défini par un couple $(\overrightarrow{dx},\overrightarrow{dy})$ porté par les axes : comme le couple $(\overrightarrow{dy},\overrightarrow{dx})$ a l'orientation opposée, les aires orientées correspondantes sont opposées. Le couple $(\overrightarrow{dx},\overrightarrow{dx})$ définit un parallélogramme aplati, donc d'aire nulle.



Avec ces règles, il vient

$$d(Pdx) = \frac{\partial P}{\partial x}dx\,dx + \frac{\partial P}{\partial y}(-dx\,dy) + \frac{\partial P}{\partial z}(dz\,dx) = -\frac{\partial P}{\partial y}dx\,dy + \frac{\partial P}{\partial z}dz\,dx$$

$$d(Qdy) = \frac{\partial Q}{\partial x}dx\,dy + \frac{\partial Q}{\partial y}dy\,dy + \frac{\partial Q}{\partial z}dz\,dy = \frac{\partial Q}{\partial x}dx\,dy - \frac{\partial Q}{\partial z}dy\,dz$$

$$d(Rdz) = \frac{\partial R}{\partial x}dx\,dz + \frac{\partial R}{\partial y}dy\,dz + \frac{\partial R}{\partial z}dz\,dz = -\frac{\partial R}{\partial x}dz\,dx + \frac{\partial R}{\partial y}dy\,dz$$

Pour la forme différentielle $\omega = Pdx + Qdy + Rdz$, on obtient $d\omega = d(Pdx) + d(Qdy) + d(Rdz)$, soit

$$d\omega = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)dy\,dz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)dz\,dx + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dx\,dy$$

Les fonctions composantes de $d\omega$ sont celles du rotationnel de $P\vec{i} + Q\vec{j} + R\vec{k}$.

Définitions

► Si A, B, C sont des fonctions, l'expression $A\,dy\,dz + B\,dz\,dx + C\,dx\,dy$ s'appelle une *forme différentielle de degré* 2. Le champ de vecteurs associé est $A\vec{i} + B\vec{j} + C\vec{k}$.

► Si ω est une forme différentielle de degré 1, de champ associé $\vec{E}$, la forme différentielle $d\omega$ de degré 2, dont le champ associé est $\overrightarrow{\text{Rot}}(\vec{E})$, s'appelle la *différentielle de* ω.

Calculons la différentielle d'une forme $\alpha = Ady\,dz+Bdz\,dx+Cdx\,dy$ de degré 2. On remplace A, B et C par leur différentielle :

$$d(Ady\,dz) = \left(\frac{\partial A}{\partial x}dx+\frac{\partial A}{\partial y}dy+\frac{\partial A}{\partial z}dz\right)dy\,dz = \frac{\partial A}{\partial x}dx\,dy\,dz+\frac{\partial A}{\partial y}dy\,dy\,dz+\frac{\partial A}{\partial z}dz\,dy\,dz$$

$$= \frac{\partial A}{\partial x}dx\,dy\,dz\,, \qquad \text{car } dy\,dy\,dz = 0 \text{ et } dz\,dy\,dz = -dy\,dz\,dz = 0$$

$$d(Bdz\,dx) = \frac{\partial B}{\partial y}dy\,dz\,dx = \frac{\partial B}{\partial y}dx\,dy\,dz\,, \qquad \text{car } dy\,dz\,dx = -dy\,dx\,dz = dx\,dy\,dz$$

$$d(Cdx\,dy) = \frac{\partial C}{\partial z}dz\,dx\,dy = \frac{\partial C}{\partial z}dx\,dy\,dz\,, \qquad \text{car } dz\,dx\,dy = -dx\,dz\,dy = dx\,dy\,dz$$

On obtient ainsi

$$d\alpha = \left(\frac{\partial A}{\partial x}+\frac{\partial B}{\partial y}+\frac{\partial C}{\partial z}\right)dx\,dy\,dz$$

La fonction entre parenthèses est la divergence du champ associé à α.

Définitions

- Une expression $f(x,y,z)\,dx\,dy\,dz$, où f est une fonction, s'appelle une *forme différentielle de degré* 3.
- Si α est une forme différentielle de degré 2, de champ associé $\vec{E}$, la forme différentielle $d\alpha = (\operatorname{div}\vec{E})\,dx\,dy\,dz$ s'appelle la *différentielle de* α.

2.2 Intégrale d'une forme différentielle

Nous savons définir l'intégrale d'une forme différentielle $\omega = Pdx+Qdy+Rdz$ sur une courbe paramétrée C : c'est la circulation $\int_C Pdx+Qdy+Rdz$ le long de C du champ associé à ω.

Intégrale d'une forme de degré 2 sur une surface

Donnons-nous une surface paramétrée régulière S : c'est l'ensemble des points $M(u,v) = \big(x(u,v), y(u,v), z(u,v)\big)$, où le paramètre (u,v) décrit un domaine Δ (page 400). Soit $\alpha = Ady\,dz+Bdz\,dx+Cdx\,dy$ une forme différentielle de degré 2, définie dans une région de l'espace contenant S.

Dans le produit $dy\,dz$, remplaçons dy et dz par leur différentielle $dy = \frac{\partial y}{\partial u}du+\frac{\partial y}{\partial v}dv$, $dz = \frac{\partial z}{\partial u}du+\frac{\partial z}{\partial v}dv$. Avec les mêmes règles $dv\,du = -du\,dv$, $du\,du = dv\,dv = 0$, il vient

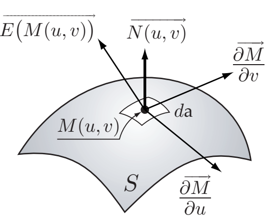
$$dy\,dz = \frac{\partial y}{\partial u}\frac{\partial z}{\partial v}du\,dv + \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}dv\,du = \left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)du\,dv$$

On calcule de même

$$dz\,dx = \left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right)du\,dv \quad , \quad dx\,dy = \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)du\,dv$$

Ainsi

$$(*)\qquad \begin{bmatrix} dy\,dz \\ dz\,dx \\ dx\,dy \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial u} \end{bmatrix} \wedge \begin{bmatrix} \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial v} \end{bmatrix} du\,dv = \left(\frac{\partial \overrightarrow{M}}{\partial u} \wedge \frac{\partial \overrightarrow{M}}{\partial v} \right) du\,dv = \overrightarrow{H(u,v)}\,du\,dv$$



Le vecteur $\overrightarrow{H(u,v)}$ est normal à S au point $M(u,v)$ et en notant $\overrightarrow{N(u,v)}$ le vecteur normal unitaire, il vient $\overrightarrow{H(u,v)}dudv = \overrightarrow{N(u,v)}\|\overrightarrow{H(u,v)}\|dudv = \overrightarrow{N(u,v)}d\text{a}$, où $d\text{a}$ est l'élément d'aire sur la surface.

Notons $\vec{E} = A\vec{i} + B\vec{j} + C\vec{k}$ le champ associé à α. Avec les paramètres u et v, la forme $\alpha = Ady\,dz + Bdz\,dx + Cdx\,dy$ s'écrit, d'après $(*)$, comme le produit scalaire

$$\overrightarrow{E(M(u,v))} \cdot \overrightarrow{H(u,v)}\,du\,dv = \overrightarrow{E(M(u,v))} \cdot \overrightarrow{N(u,v)}\,d\text{a}$$

Définition

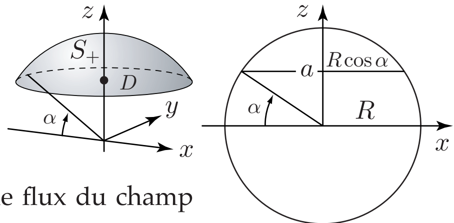
Soit $\alpha = Ady\,dz + Bdz\,dx + Cdx\,dy$ une forme différentielle de degré 2 et soit $\vec{E} = A\vec{i} + B\vec{j} + C\vec{k}$ son champ associé. L'intégrale de α sur S est

$$\int_S \alpha = \iint_\Delta \overrightarrow{E(M(u,v))} \cdot \overrightarrow{N(u,v)}\ d\text{a}\,,$$

où $\overrightarrow{N(u,v)}$ est le vecteur normal à S et $d\text{a}$ l'élément d'aire sur la surface.

Si par exemple $\vec{E}$ est un champ de vitesses de particules, l'intégrale de α sur S est la quantité de matière qui traverse la surface par unité de temps. C'est pourquoi cette intégrale s'appelle aussi le *flux de $\vec{E}$ à travers S*.

Exemple. Coupons la sphère de rayon R centrée à l'origine par le plan horizontal d'équation $z = a$, où $0 < a < R$. Considérons la surface fermée S formée de la calotte supérieure de la sphère et du disque horizontal situé dans le plan de coupe. Calculons le flux du champ $\vec{E} = (1/z)\vec{k}$ à travers S, c'est-à-dire l'intégrale sur S de la forme différentielle $\frac{1}{z}\,dxdy$. Notons S_+ la calotte et D le disque qui lui sert de base.



Flux à travers S_+. On paramètre la sphère par les coordonnées sphériques $x = R\cos\theta\cos\varphi$, $y = R\sin\theta\cos\varphi$, $z = R\sin\varphi$, où $0 \leqslant \theta < 2\pi$ et $-\pi/2 \leqslant \varphi \leqslant \pi/2$. Le vecteur normal à la sphère est alors dirigé vers l'extérieur. En posant $a = R\sin\alpha$, avec $0 < \alpha < \pi/2$, la calotte est formée des points tels que $0 \leqslant \theta < 2\pi$ et $\alpha \leqslant \varphi \leqslant \pi/2$. On a

$$dx\,dy = d(R\cos\theta\cos\varphi)\,d(R\sin\theta\cos\varphi) = R^2\sin\varphi\cos\varphi\,d\theta\,d\varphi$$

$$\iint_{S_+} \frac{dx\,dy}{z} = \int_\alpha^{\pi/2}\int_0^{2\pi} \frac{R^2\sin\varphi\cos\varphi}{R\sin\varphi}\,d\theta d\varphi = 2\pi R\int_\alpha^{\pi/2}\cos\varphi\,d\varphi = 2\pi R(1-\sin\alpha)$$

Flux à travers D. En tout point de D, le champ est $(1/a)\vec{k}$, un vecteur constant. Le vecteur normal à D orienté vers l'extérieur est $\vec{N} = -\vec{k}$, donc le flux de $\vec{E}$ à travers D est le produit de l'aire de D par $(1/a)\,\vec{k}\cdot\vec{N} = -1/a$. Puisque le rayon de D est $R\cos\alpha$, ce flux est $\int_D \dfrac{dx\,dy}{z} = \pi(R\cos\alpha)^2\left(\dfrac{-1}{a}\right) = -\pi R\dfrac{(\cos\alpha)^2}{\sin\alpha}$.

Le flux de $\vec{E}$ à travers S est donc

$$2\pi R(1-\sin\alpha) - \pi R\frac{(\cos\alpha)^2}{\sin\alpha} = 2\pi R - \frac{\pi R}{\sin\alpha}\left[1+(\sin\alpha)^2\right] = -\pi R\frac{(1-\sin\alpha)^2}{\sin\alpha}.$$

2.3 Formule de Stokes

Nous allons considérer des domaines à bord convenablement paramétrés.

- Des surfaces régulières dont le bord est réunion de courbes régulières ; ce sont des domaines de *dimension* 2, leur bord est de dimension 1.
  Exemples : un rectangle plein, un disque, une calotte sphérique ou la surface latérale d'un cylindre.
- Des domaines de l'espace limités par des morceaux de surfaces régulières qui en constituent le bord ; ces domaines sont de *dimension* 3 et leur bord est de dimension 2.
  Exemples : une sphère pleine ou un cylindre plein, l'intérieur d'un parallélépipède ou un tore plein.

Orientation et règles de paramétrage. Rappelons la définition du vecteur normal unitaire à une courbe ou à une surface paramétrée.

- En tout point $M(t)$ d'une courbe paramétrée plane régulière, le vecteur tangent unitaire $\vec{T}$, porté par $\dfrac{\partial\overrightarrow{M}}{\partial t}$, et le vecteur normal unitaire $\vec{N}$, forment une base orthonormée directe $(\vec{T},\vec{N})$ (page 315).
- En tout point $M(u,v)$ d'une surface paramétrée régulière, le vecteur normal unitaire est dans la direction du produit vectoriel $\dfrac{\partial\overrightarrow{M}}{\partial u} \wedge \dfrac{\partial\overrightarrow{M}}{\partial v}$ (page 400).

Règle 1. Une courbe C formant le bord d'une surface S sera toujours paramétrée de manière que le vecteur normal à C et tangent à S soit dirigé vers *l'intérieur* du domaine : pour une surface plane, cela veut dire que si l'on parcourt le bord dans le sens croissant du paramètre, alors l'intérieur du domaine est du côté gauche.

Règle 2. Pour une région de l'espace limitée par une surface, on doit paramétrer cette surface de manière que le vecteur normal soit toujours dirigé vers *l'extérieur* du domaine.

Notation. Si D est un domaine, on note ∂D son bord ainsi paramétré.

Exemples. (voir les figures ci-dessous)

1) Le domaine D_1 est un disque centré à l'origine, son bord ∂D_1 est un cercle orienté dans le sens de la flèche (figure 1). Si le rayon est R, on peut prendre comme paramétrage du bord $M(t) = \big(x(t), y(t)\big) = (R\cos t, R\sin t)$, où t va de 0 à 2π : le vecteur normal en (x, y) est $\dfrac{1}{R}(-y, x)$, il est bien dirigé vers l'intérieur du disque.

2) D_2 est le domaine compris entre deux disques et son bord ∂D_2 est constitué de deux cercles (figure 2) ; ceux-ci sont orientés dans des sens différents (le sens des flèches est celui qui met le domaine à gauche quand on se déplace sur le bord). On paramètre le cercle extérieur comme ci-dessus. Si le cercle intérieur, de rayon $r < R$, est centré en (a, b), on peut le paramétrer par $x = a + r\cos t$, $y = b - r\sin t$, t allant de 0 à 2π : la normale pointe alors vers l'intérieur du domaine D_2.

3) Le domaine D_3 est un hémisphère, formé des points $(x, y, \sqrt{R^2 - x^2 - y^2})$, où R est le rayon (figure 3). Son bord est le grand cercle dans le plan xOy, orienté comme l'indique la flèche.

4) Le domaine $V = \{(x, y, z) \mid x^2 + y^2 + z^2 \leqslant R^2\}$ est une boule (pleine) de rayon R et son bord $S = \partial V$ est une sphère (figure 4). En utilisant les coordonnées sphériques, les points de S sont

$$M(\theta, \varphi) = \big(R\cos\theta\cos\varphi, R\sin\theta\cos\varphi, R\sin\varphi\big), \quad \text{où } 0 \leqslant \theta < 2\pi \text{ et } -\pi/2 \leqslant \varphi \leqslant \pi/2.$$

Le vecteur normal en M est $\dfrac{\partial\overrightarrow{M}}{\partial\theta} \wedge \dfrac{\partial\overrightarrow{M}}{\partial\varphi} = (R\cos\varphi)\overrightarrow{OM}$, donc il est dirigé vers l'extérieur de la boule, puisque $\cos\varphi > 0$.

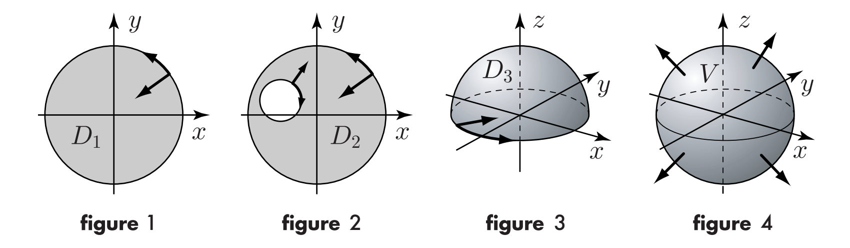


figure 1 **figure 2** **figure 3** **figure 4**

Formule de Stokes. *Soit D un domaine comme ci-dessus et soit ω une forme différentielle définie dans une région contenant D et de degré la dimension du bord de D. Alors on a*

$$\int_{\partial D} \omega = \int_D d\omega$$

Voyons comment se traduit cette formule selon les cas.

Cas d'un domaine plan

Supposons que D est un domaine plan limité par une courbe fermée C. Prenons une forme différentielle à deux variables : $\omega = P dx + Q dy$, définie dans une région

contenant D. On a alors $d\omega = \left(\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y}\right) dx\, dy$ et la formule de Stokes s'écrit

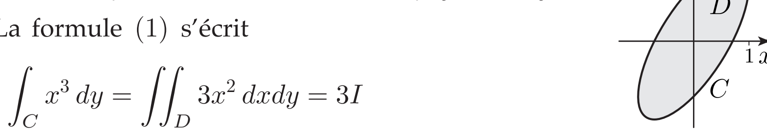
$$(1) \qquad \int_C P dx + Q dy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

Justification. Prenons comme domaine D un rectangle, formé des points (x,y) tels que $a \leqslant x \leqslant b$ et $p \leqslant y \leqslant q$. Le bord est constitué de deux segments horizontaux h_+ et h_- et de deux segments verticaux v_+ et v_-. Calculons l'intégrale $I = \iint_D -\dfrac{\partial P}{\partial y}\, dxdy$ en intégrant d'abord pour x fixé :

$$I = \int_a^b \left(\int_p^q -\frac{\partial P}{\partial y}\, dy\right) dx = \int_a^b \left[-P(x,q) + P(x,p)\right] dx$$

En paramétrant h_- par $(x = t, y = p)$, où t parcourt $[a, b]$, il vient $\int_{h_-} P\, dx = \int_a^b P(t,p)\, dt$. Pour l'intégrale curviligne sur h_+, l'abscisse du point doit aller de b à a, donc $\int_{h_+} P\, dx = \int_b^a P(t,q)\, dt = -\int_a^b P(t,q)\, dt$. On a donc $I = \int_{h_+ \cup h_-} P\, dx$. Sur les segments verticaux, x est constant, la différentielle dx est nulle, donc $\int_{v_+} P\, dx = \int_{v_-} P\, dx = 0$ et finalement $I = \int_{\partial D} P\, dx$. De même, on a $\iint_D \dfrac{\partial Q}{\partial x}\, dxdy = \int_{\partial D} Q\, dy$ et il vient $\iint_D \left(\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y}\right) dxdy = \int_{\partial D} P dx + Q dy$. ■

Exemple. Calculons l'intégrale $I = \iint_D x^2\, dxdy$, où D est l'intérieur de l'ellipse C d'équation $(y - x)^2 + x^2 = 1$.



En choisissant $P = 0$ et $Q = x^3$, il vient $P dx + Q dy = x^3\, dy$ et $\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y} = 3x^2$. La formule (1) s'écrit

$$\int_C x^3\, dy = \iint_D 3x^2\, dxdy = 3I$$

Paramétrons l'ellipse en posant $x = \cos t$ et $y - x = \sin t$, donc $y = \sin t + \cos t$. On a $dy = \cos t - \sin t$ et $\int_C x^3\, dy = \int_{-\pi}^{\pi} (\cos t)^3 (\cos t - \sin t)\, dt = \int_{-\pi}^{\pi} (\cos t)^4\, dt$, car la fonction $(\cos t)^3 \sin t$ étant impaire, son intégrale entre $-\pi$ et π est nulle. On obtient ainsi $3I = 3\pi/4$, d'où $I = \pi/4$. Si D est une plaque homogène de densité ρ, son moment d'inertie par rapport à l'axe Oy est donc $\rho\pi/4$.

Cas d'une surface dans l'espace

Supposons que le domaine est une surface S dans l'espace, bordée par une courbe C.

- La forme différentielle est $\omega = P dx + Q dy + R dz$, son champ est $\vec{E} = P\vec{i} + Q\vec{j} + R\vec{k}$ et l'intégrale de ω sur $C = \partial S$ est la circulation de $\vec{E}$ sur C.
- Le champ associé à $d\omega$ est $\overrightarrow{\text{Rot}}(\vec{E})$ et l'intégrale de $d\omega$ sur S est le flux de ce rotationnel.

La formule de Stokes s'écrit

$$\int_C P dx + Q dy + R dz = \iint_S \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dydz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dzdx + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

$$(2) \qquad \int_C Pdx+Qdy+Rdz = \iint_S \overrightarrow{\mathrm{Rot}}(\vec{E}) \cdot \vec{N}\, d\mathrm{a}$$

Si un champ de vecteur $\vec{E}$ est défini dans une région contenant la surface S, la circulation de $\vec{E}$ sur le bord de S est le flux du rotationnel de $\vec{E}$ à travers S.

Cas d'un domaine de l'espace limité par une surface fermée

Soient S une surface fermée et D la région de l'espace située à l'intérieur.

La forme différentielle est $\alpha = Adydz + Bdzdx + Cdxdy$, son champ est $\vec{E} = A\vec{i} + B\vec{j} + C\vec{k}$.

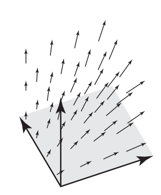
L'intégrale de α sur $S = \partial D$ est le flux de $\vec{E}$ à travers S.

Puisque $d\alpha = (\mathrm{div}\,\vec{E})\,dxdydz$, la formule de Stokes s'écrit

$$(3) \qquad \iiint_D (\mathrm{div}\,\vec{E})\,dxdydz = \iint_S \vec{E} \cdot \vec{N}\, d\mathrm{a}$$

Si un champ de vecteur E est défini dans une région de l'espace D bordée par une surface fermée S, le flux de $\vec{E}$ à travers S est l'intégrale dans D de la divergence de $\vec{E}$.

Si la divergence de $\vec{E}$ est nulle, le flux à travers toute surface fermée est nul : on dit que le champ est *conservatif*. Un champ de vitesses est conservatif si dans toute région de l'espace, la quantité de matière entrante par unité de temps est égale à la quantité sortante (exercice 6). On a représenté ci-contre le champ (x, y, e^{z^2}) : il n'est pas conservatif.



Exemple. Reprenons la surface fermée S de l'exemple page 428. D'après la formule (3), le flux du champ $\vec{E} = (1/z)\vec{k}$ à travers S est l'intégrale de $\mathrm{div}\,\vec{E} = -1/z^2$ dans l'intérieur Δ de S. Pour calculer directement cette intégrale, intégrons d'abord pour z fixé entre a et R : la section de Δ par le plan horizontal à la hauteur z est un disque de rayon $r(z) = \sqrt{R^2 - z^2}$, donc d'aire $\pi(R^2 - z^2)$. En intégrant d'abord en x puis en y, il vient

$$\iiint_\Delta (\mathrm{div}\,\vec{E})\,dxdydz = \iiint_\Delta \frac{-1}{z^2}\,dxdydz = \int_a^R \pi(R^2 - z^2)\frac{-1}{z^2}\,dz = \pi(R-a) + \pi R^2\left(\frac{1}{R} - \frac{1}{a}\right)$$

Puisque $a = R\sin\alpha$, cela s'écrit $\pi R(1 - \sin\alpha) + \pi R\left(1 - \frac{1}{\sin\alpha}\right) = \pi R\left(2 - \sin\alpha - \frac{1}{\sin\alpha}\right)$ ou encore

$$\iiint_\Delta (\mathrm{div}\,\vec{E})\,dxdydz = -\pi R\frac{(1 - \sin\alpha)^2}{\sin\alpha}$$

et l'on retrouve bien la valeur du flux de $\vec{E}$ à travers S.

2.4 Applications

Propriété de l'extremum pour une fonction harmonique

Soit f une fonction harmonique dans une région R de l'espace, région qu'on suppose d'un seul tenant. Nous allons voir qu'à moins d'être constante, f ne peut avoir ni minimum, ni maximum en un point intérieur à R.

Supposons que f a un minimum en un point M_0 intérieur à R. Soit S une petite surface de niveau entourant M_0 et soit D le domaine intérieur à S. Puisque le gradient de f pointe vers les niveaux supérieurs, on sait qu'en tout point M, le vecteur $\overrightarrow{E(M)} = \overrightarrow{\mathrm{Grad}}_f(M)$ pointe vers l'extérieur de D : cela veut dire que si $\overrightarrow{N(M)}$ est le vecteur normal à S en M, le produit scalaire $\overrightarrow{E(M)} \cdot \overrightarrow{N(M)}$ est positif ou nul, et par suite

$$\iint_S \overrightarrow{E(M)} \cdot \overrightarrow{N(M)}\, d\mathrm{a} \geqslant 0$$

D'après la formule 3, l'intégrale ci-dessus vaut

$$I = \iiint_D \mathrm{div}\left(\overrightarrow{\mathrm{Grad}}_f\right)\, dxdydz = \iiint_D \Delta f\ dxdydz = 0$$

car le laplacien Δf est nul. On en déduit qu'en tout point $M \in S$, on a $\overrightarrow{E(M)} \cdot \overrightarrow{N(M)} = 0$, autrement dit $\overrightarrow{E(M)}$ est tangent à S : au voisinage de M, il n'y a pas de niveau supérieur à celui de M, donc $\overrightarrow{E(M)} = \overrightarrow{\mathrm{Grad}}_f(M) = 0$. Cela étant vrai pour tous les points M voisins de M_0, on en déduit que f est constante dans D.
Considérons maintenant un point M_1 quelconque dans D. Par hypothèse, on peut joindre M_0 à M_1 par une courbe M_t tracée dans R, où t parcourt (par exemple) $[0,1]$. Pour t assez proche de 0, M_t est dans D, donc $f(M_t) = f(M_0)$. Soit τ la plus grande valeur de t telle que $f(M_t) = f(M_0)$. Puisque $f(M_0)$ est la valeur minimum de f sur R, le raisonnement précédent montre que f est constante au voisinage du point M_τ. Si $\tau < 1$, on a donc aussi $f(M_t) = f(M_\tau) = f(M_0)$ pour des valeurs de t un peu supérieures à τ : c'est impossible, d'après le choix de τ. C'est donc qu'on a $\tau = 1$ et $f(M_1) = f(M_\tau) = f(M_0)$. La fonction f est donc constante sur R.

Recherche d'un potentiel harmonique. Nous avons montré que dans un fluide incompressible en mouvement non tourbillonnaire, le champ des vitesses est de la forme $-\overrightarrow{\mathrm{Grad}}_\rho$, où ρ est une fonction harmonique à l'intérieur du domaine R occupé par le fluide. En général, on connaît les valeurs de ρ sur le bord de R et éventuellement la limite de $\rho(M)$ quand M tend vers l'infini : c'est ce qu'on appelle les conditions au bord.

Supposons que ρ_1 et ρ_2 sont des potentiels harmoniques à l'intérieur de R satisfaisant les mêmes conditions au bord. La fonction $f = \rho_2 - \rho_1$ est harmonique dans R et comme elle est nulle au bord, f atteint un extremum en un point intérieur à R. D'après ce qui précède, f est constante dans R, et comme f est nulle au bord, on a $f(M) = 0$ en tout point M de R, c'est-à-dire $\rho_1 = \rho_2$.

Le potentiel harmonique ρ est donc parfaitement déterminé par les conditions aux bord du domaine, et il en va de même du mouvement des particules dans le fluide.

Calcul d'une aire plane

Soit D un domaine plan limité par une courbe C. Dans la formule de Stokes (1), prenons $P(x,y) = -y$ et $Q = 0$. Alors $\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y} = 1$ et il vient

$$\text{Aire de } D = \int_C -y\,dx$$

De même, avec $Q(x,y) = x$ et $P = 0$, on a : Aire de $D = \int_C x\,dy$. Il est parfois commode de faire intervenir les deux coordonnées x et y et d'utiliser la formule

$$\text{Aire de } D = \frac{1}{2}\int_C x\,dy - y\,dx$$

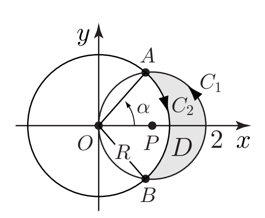
Aire d'un domaine fermé paramétré par les coordonnées polaires

Prenons dans le plan une courbe fermée C formée des points M dont les coordonnées polaires (r,θ) vérifient $r = f(\theta)$, où f est une fonction dérivable sur un intervalle contenu dans $[0,2\pi]$. Utilisons la formule précédente pour calculer l'aire du secteur bordé par C, un résultat qu'on obtient aussi par calcul direct en coordonnées polaires (exercice 1 page 411).
En tout point de C, on a

$$x = r\cos\theta = f(\theta)\cos\theta \quad , \quad dx = f'(\theta)\cos\theta - f(\theta)\sin\theta$$
$$y = r\sin\theta = f(\theta)\sin\theta \quad , \quad dy = f'(\theta)\sin\theta + f(\theta)\cos\theta$$

d'où $x\,dy - y\,dx = \Big([f(\theta)]^2(\cos\theta)^2 + [f(\theta)]^2(\sin\theta)^2\Big)d\theta = \big[f(\theta)\big]^2 d\theta$ et l'aire de D est $\dfrac{1}{2}\int_C \big[f(\theta)\big]^2 d\theta$.

Exemple. Dans le disque de rayon 1 centré au point $P{=}(1{,}0)$, enlevons la partie incluse dans un disque de rayon $R<2$ centré à l'origine O : on obtient un domaine D en forme de croissant. Notons A la pointe supérieure, B la pointe inférieure et α l'angle $\widehat{\overrightarrow{OP},\overrightarrow{OA}}$.



Le bord de D est formé de deux arcs de cercles :

- l'arc C_1 du cercle de centre P, d'équation $r = 2\cos\theta$, $-\alpha \leqslant \theta \leqslant \alpha$,
- l'arc C_2 du cercle de centre O, d'équation $r = R$, $-\alpha \leqslant \theta \leqslant \alpha$.

Comme les points A et B sont à la distance R de l'origine, l'angle α est déterminé par l'égalité $R = 2\cos\alpha$. Paramétrons ce bord dans le sens des flèches indiquées sur la figure. Il vient

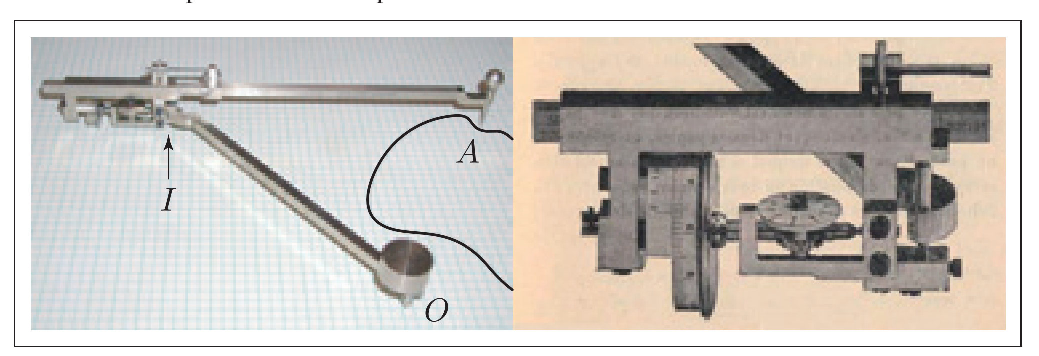
$$\frac{1}{2}\int_{C_1} r^2\,d\theta = \int_{-\alpha}^{\alpha} 2(\cos\theta)^2\,d\theta = \int_{-\alpha}^{\alpha}\big(1+\cos 2\theta\big)\,d\theta = 2\alpha + \sin 2\alpha$$

$$\frac{1}{2}\int_{C_2} r^2\,d\theta = \frac{1}{2}\int_{\alpha}^{-\alpha} R^2\,d\theta = -\alpha R^2 = -4\alpha(\cos\alpha)^2$$

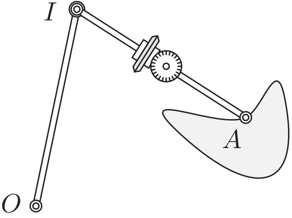
L'aire de D est la somme de ces deux intégrales, c'est-à-dire $2\alpha\left[1-2(\cos\alpha)^2\right]+\sin 2\alpha = \sin 2\alpha - 2\alpha\cos 2\alpha$.

2.5 Le planimètre

Le planimètre, inventé par J. Amsler en 1854, est un instrument qui permet de mesurer l'aire plane délimitée par une courbe fermée.



Description. L'appareil est constitué de deux tiges OI et IA de même longueur et articulées en I.



- L'extrémité O peut être fixée en un point d'un support plan, tandis que l'extrémité A peut bouger librement sur ce plan.
- Sur la tige IA est fixée une roulette d'axe IA ; on la munit d'un compte-tour, car elle est destinée à enregistrer les mouvements de rotation de la tige.

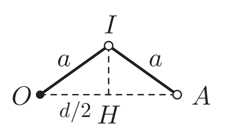
Fonctionnement. On fixe O en un point du plan de la courbe et l'on fait parcourir la courbe à l'extrémité A de la tige libre. Pendant le mouvement, la roulette tourne par frottement sur le support.

Le nombre de tours de la roulette est proportionnel à l'aire intérieure à la courbe.

Il suffit donc d'étalonner l'instrument sur un carré unité pour avoir un outil qui mesure les aires.

Explication. Pour repérer les points du plan, prenons des axes orthonormés d'origine O. Calculons les coordonnées (p,q) de I en fonction des coordonnées (x,y) de A.

Notons a la longueur commune des tiges OI et IA. Le triangle OIA étant isocèle, la hauteur IH issue de I est une médiane. En posant $d=OA=2OH$, il vient donc (théorème de Pythagore) $HI^2 = OI^2 - OH^2 = a^2 - \dfrac{d^2}{4}$. Les coordonnées du vecteur unitaire porté



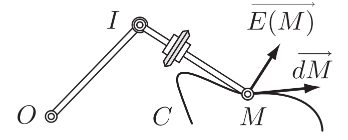
par $\overrightarrow{OA}$ sont $(x/d, y/d)$, donc $\overrightarrow{HI}$ a pour coordonnées $\sqrt{a^2-(d^2/4)}(-y/d, x/d)$. Les coordonnées de $\overrightarrow{OH}$ sont $(x/2, y/2)$ et $\overrightarrow{OI}=\overrightarrow{OH}+\overrightarrow{HI}$, donc les coordonnées du point I sont

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} x/2 \\ y/2 \end{bmatrix} + \frac{1}{d}\sqrt{a^2 - \frac{d^2}{4}} \begin{bmatrix} -y \\ x \end{bmatrix}, \quad \text{avec } d = \sqrt{x^2+y^2}.$$

$$\text{Coordonnées de } I : \quad p = \frac{x}{2} - \frac{y}{2}\sqrt{\frac{4a^2-(x^2+y^2)}{x^2+y^2}} \quad , \quad q = \frac{y}{2} + \frac{x}{2}\sqrt{\frac{4a^2-(x^2+y^2)}{x^2+y^2}}.$$

Si M est un point du plan accessible au planimètre, amenons A en ce point et notons $\overrightarrow{E(M)}$ le vecteur unitaire directement orthogonal à $\overrightarrow{IM}$. Les coordonnées de $\overrightarrow{IM}$ étant $(x-p, y-q)$, celles de $\overrightarrow{E(M)}$ sont $(P, Q) = \frac{1}{a}\big(-(y-q)\ ,\ x-p\big)$.
On définit ainsi un champ de vecteurs unitaires dans une région du plan.

Premier point. *Le nombre de tours de roulette est proportionnel à la circulation du champ $\overrightarrow{E(M)}$ le long de la courbe C.*



Si l'on déplace M dans l'axe de la roulette, elle ne tourne pas. Si l'on déplace M dans le sens de $\overrightarrow{E(M)}$, elle tourne d'une quantité proportionnelle au déplacement. Pour un petit déplacement $\overrightarrow{dM}$ de direction quelconque, la rotation de la roulette est proportionnelle à la composante de $\overrightarrow{dM}$ sur $\overrightarrow{E(M)}$, c'est-à-dire au produit scalaire $\overrightarrow{E(M)} \cdot \overrightarrow{dM}$. Quand M parcourt la courbe, l'accroissement $\overrightarrow{dM}$ reste tangent à C et le nombre de tours de roulette est donc proportionnel à $\int_C \overrightarrow{E(M)} \cdot \overrightarrow{dM}$.

Deuxième point. *En tout point M, on a $\overrightarrow{\mathrm{Rot}}(\vec{E}) = \frac{1}{a}\vec{k}$, où $\vec{k}$ est un vecteur unitaire orthogonal au plan.*

En posant $d = x^2+y^2$ et $u = \sqrt{\frac{4a^2-d^2}{d^2}}$, on a en effet $\frac{\partial Q}{\partial x} = \frac{1}{2a} - \frac{2axy}{ud^4}$ et $\frac{\partial P}{\partial y} = \frac{-1}{2a} - \frac{2axy}{ud^4}$, donc $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = \frac{1}{a}$.

Troisième point. Soit D l'intérieur de la courbe. D'après la formule de Stokes, la circulation du champ $\overrightarrow{E(M)}$ le long de C est

$$\int_C \overrightarrow{E(M)} \cdot \overrightarrow{dM} = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy = \iint_D \frac{1}{a}\, dxdy = \frac{1}{a} \text{ aire de } D.$$

On en déduit que l'aire intérieure à C est proportionnelle au nombre de tours enregistrés par la roulette quand M parcourt la courbe.

Exercices

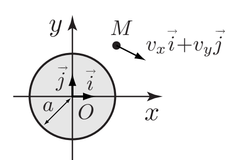
@ **1. Gradients conjugués.** Soit $\vec{E} = P\vec{i} + Q\vec{j}$ un champ de vecteurs défini dans le plan. Définissons le champ de vecteurs $\vec{E^c} = -Q\vec{i} + P\vec{j}$.

a) Montrer qu'en tout point M du plan, le vecteur $\overrightarrow{E^c(M)}$ se déduit de $\overrightarrow{E(M)}$ par rotation d'angle $+\frac{\pi}{2}$.

b) On suppose que $\vec{E} = \overrightarrow{\mathrm{Grad}}_\varphi$, où φ est une fonction harmonique. Montrer que $\overrightarrow{\mathrm{Rot}}\left(\vec{E^c}\right) = 0$. En déduire qu'il existe une fonction φ^c telle que $\vec{E^c} = \overrightarrow{\mathrm{Grad}}_{\varphi^c}$ et que φ^c est harmonique. On dit que φ et φ^c sont des *potentiels conjugués*.

c) On prend $\vec{E} = (x^2-y^2)\vec{i} - 2xy\vec{j}$. Montrer que le champ E est le gradient d'une fonction φ. Calculer $\varphi(x,y)$ et le potentiel conjugué $\varphi^c(x,y)$.

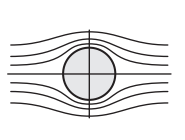
@ **2. Un modèle d'écoulement.** On considère le mouvement plan non tourbillonnaire d'un fluide autour d'un cylindre fixe de rayon a et d'axe vertical. Plaçons l'origine O des coordonnées sur l'axe du cylindre, Ox étant l'axe de symétrie du mouvement. En appelant v_x,v_y les composantes de la vitesse d'une particule dans le fluide et r la distance à l'origine, le champ des vitesses est $v_x\vec{i} + v_y\vec{j} = \overrightarrow{\mathrm{Grad}}_\varphi$, où $\varphi = x + \dfrac{a^2x}{r^2}$.

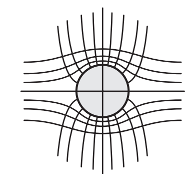


a) Écrire $\varphi(x,y)$ au moyen de r et θ. En déduire que la fonction φ est harmonique (se reporter page 372). Montrer que $v_x = 1 - a^2\dfrac{y^2-x^2}{r^4}$ et $v_y = -a^2\dfrac{2xy}{r^4}$.

b) Montrer que le potentiel conjugué (voir l'exercice ci-dessus) est $\varphi^c = y - \dfrac{a^2y}{r^2}$. En déduire que les lignes de courant sont les courbes d'équation $\varphi^c(x,y) = K$, où K est une constante.

c) À l'aide d'un ordinateur, vérifier que les lignes de courant ont l'allure montré figure **1**. Sur la figure **2**, on voit aussi les courbes d'équation $\varphi =$ constante : ce sont les lignes d'égale pression. En chaque point, ligne de courant et ligne d'égale pression sont orthogonales.


figure 1


figure 2

En réalité, cette modélisation n'est pas correcte dans la région juste en aval du cylindre : on y observe un effet de sillage plus ou moins important selon la viscosité et la vitesse ; aux vitesses élevées, il se forme des tourbillons.

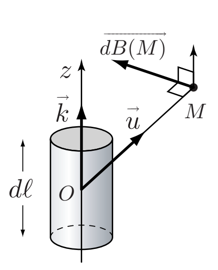
3. Le champ électrique. Le champ électrique créé en un point M de l'espace par une charge q placée à l'origine O des coordonnées est $\overrightarrow{E(M)} = \dfrac{q}{4\pi\varepsilon_0}\overrightarrow{F(M)}$, où $\overrightarrow{F(M)} = \dfrac{1}{r^2}\vec{u}$, $\vec{u}$ étant le vecteur unitaire dirigé par $\overrightarrow{OM}$ et r la distance OM.

a) Calculer la divergence $\operatorname{div}\vec{F} = \dfrac{\partial}{\partial x}\left(\dfrac{x}{r^3}\right) + \dfrac{\partial}{\partial y}\left(\dfrac{y}{r^3}\right) + \dfrac{\partial}{\partial z}\left(\dfrac{z}{r^3}\right)$. En déduire que le flux de $\vec{E}$ à travers une surface fermée n'entourant pas la charge est nul.

b) Soit S la sphère centrée à l'origine et de rayon a. Montrer qu'en tout point M de S, le vecteur normal unitaire dirigé vers l'extérieur est $\vec{N} = \dfrac{\overrightarrow{OM}}{a}$ et que $\vec{F}\cdot\vec{N} = \dfrac{1}{a^2}$. En se rappelant que l'aire de S est $4\pi a^2$, en déduire que le flux de $\vec{E}$ à travers S est $\displaystyle\int_S \vec{E}\cdot\vec{N}\,\mathrm{d}a = \frac{q}{\varepsilon_0}$.

Le flux du champ électrique à travers une sphère centrée sur la charge ne dépend donc pas du rayon de la sphère : c'est la loi de Gauss en électrostatique.

4. Le champ magnétique. Un « élément de courant électrique », formé d'un fil rectiligne de longueur infinitésimale $d\ell$ placé en un point O et parcouru par un courant d'intensité j, crée en tout point M un champ magnétique $\overrightarrow{dB(M)} = \dfrac{\mu_0}{4\pi}\dfrac{j\,d\ell\;\vec{k}\wedge\vec{u}}{OM^3}$, où $\vec{k}$ est le vecteur unitaire dans la direction du fil dirigé dans le sens du courant et $\vec{u}$ le vecteur unitaire dans la direction $\overrightarrow{OM}$. On choisit l'origine des coordonnées en O et l'axe Oz dirigé selon $\vec{k}$.



Calculer les composantes B_x, B_y, B_z du champ $\overrightarrow{dB(M)}$ et montrer que la divergence de $\overrightarrow{dB}$ est nulle.

Le champ magnétique créé par un courant est conservatif.

@ **5.** Déterminer la fonction $y \mapsto f(y)$ pour que le champ $e^x f(y)\,\vec{i} + e^x \ln y\,\vec{j}$ dérive d'un potentiel. Quel est le potentiel qui tend vers 0 quand (x,y) tend vers $(1,0)$ et vers 1 quand (x,y) tend vers $(0,0)$? On trouve $f(y) = y\ln y - y + c$, où c est une constante et le potentiel est $\rho = -\left[e^x(y\ln y - y + c) + d\right]$, avec $c = \dfrac{1}{e-1}$ et $d = \dfrac{-e}{e-1}$.

6. Soit $\vec{E}$ un champ de vecteurs tel que $\operatorname{div}\vec{E} = 0$. Considérons un tube de champ basé sur une surface s : il est constitué des lignes de champ aa' issues des points de s ; limitons ce tube par la surface s'.
Montrer que le flux de $\vec{E}$ à travers la surface latérale du tube est nulle. En déduire que le flux de $\vec{E}$ à travers une section quelconque du tube ne dépend pas de cette section (flux conservatif).
Montrer que si le tube est très étroit et les sections s et s' orthogonales aux lignes de champ, les aires de s et s' sont inversement proportionnelles aux intensités $\|\vec{E}\|$ et $\|\vec{E'}\|$ du champ sur s et s'.

Chapitre 15

Équations différentielles

Une *équation différentielle* est une relation entre une fonction $x(t)$ et ses dérivées $x'(t), x''(t), \ldots$. Dans le cas le plus simple, elle permet d'exprimer pour tout t, la valeur de $x'(t)$ au moyen de $x(t)$ et de t, sous la forme $x'(t) = f\big(t, x(t)\big)$, où f est une fonction de deux variables. Une telle équation s'écrit simplement $x' = f(t, x)$.

Exemples

1) $x' = -2tx$ est une équation différentielle, du premier ordre car elle ne fait intervenir que la dérivée première. Une solution est une fonction $u(t)$ dérivable telle que $u'(t) = -2tu(t)$ pour tout t.

   La fonction $u(t) = \exp(-t^2)$ est solution, car $u'(t) = -2t\exp(-t^2) = -2tu(t)$. De même, pour tout nombre λ, la fonction $\lambda \exp(-t^2)$ est solution.

2) $x' = x$ est aussi une équation différentielle du premier ordre. Ses solutions sont les fonctions dérivables $u(t)$ telles que $u'(t) = u(t)$ pour tout t : la fonction $u(t) = \exp(t)$ est une solution, de même que $\lambda \exp(t)$ pour tout nombre λ.

3) $x'' = x' + 2x - 2\exp t$ est une équation différentielle du second ordre, dont les solutions sont les fonctions $u(t)$ (deux fois dérivables) vérifiant $u''(t) = u'(t) + 2u(t) - 2\exp t$ pour tout t.

   Pour tous nombres λ et μ, les fonctions $\lambda\exp(2t) + \mu\exp(-t) + \exp t$ sont des solutions de cette équation.

4) Si ω est un nombre donné, la fonction $u(t) = \sin \omega t$ est solution de l'équation différentielle $x'' + \omega^2 x = 0$, car on a $u'(t) = \omega \cos \omega t$ et $u''(t) = -\omega^2 \cos \omega t = -\omega^2 u(t)$; de même, $v(t) = \cos \omega t$ est une solution.

   Les solutions sont toutes les fonctions $\lambda \cos \omega t + \mu \sin \omega t$, où λ et μ sont des nombres quelconques.

1. Équations différentielles du premier ordre

1.1 L'approche graphique

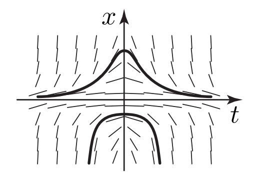
Soit $x' = f(t,x)$ une équation différentielle du premier ordre, la fonction f étant définie dans un certain domaine D du plan.

Supposons que u est une solution et soit $M = (\tau, b)$ un point situé sur le graphe de u. On a $b = u(\tau)$, donc $u'(\tau) = f(\tau, b)$ est un nombre qui ne dépend que du point M. Puisque la tangente en M au graphe de u a pour pente $u'(\tau)$, on en déduit les propriétés suivantes.

- *En tout point (t,x) du graphe d'une solution, la tangente a pour pente $f(t,x)$.*
- *Dans une région où $f(t,x) > 0$, les solutions sont croissantes ; dans une région où $f(t,x) < 0$, elles sont décroissantes.*

Voici comment on peut deviner l'allure des solutions. En tout point $M = (t,x)$ de D, dessinons un petit segment centré en M et de pente $f(t,x)$: on obtient un *champ de directions*. Par définition, une solution de l'équation différentielle est une fonction dont le graphe est tangent en tout point aux segments ainsi dessinés.

Exemple. La figure ci-contre montre le champ de directions pour l'équation différentielle $x' = -2tx^2$: en tout point (t,x), nous avons placé un segment de pente $f(t,x) = -2tx^2$. Ainsi, sur l'axe des ordonnées ($t = 0$), les segments sont tous horizontaux et il en va de même sur l'axe des abscisses ($x = 0$). Il suffit de relier les segments pour avoir l'allure de quelques solutions. On voit que la fonction nulle, dont le graphe est l'axe des abscisses, est solution. Puisque $-2tx^2$ a le signe de $-t$, les solutions sont décroissantes pour $t > 0$ et croissantes pour $t < 0$.



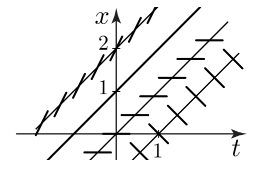
En deux points (t,x), $(-t,x)$ symétriques par rapport à Ox, les pentes du champ de directions sont opposées, car on a $f(-t,x) = -f(t,x)$: en conséquence, si une solution est définie sur un intervalle de la forme $]-a,a[$, c'est une fonction paire.

Les isoclines

Si p est un nombre donné, les points (t,x) où les solutions ont une tangente de pente p vérifient $f(t,x)=p$. C'est l'équation d'une courbe, appelée *l'isocline pour la pente p*, car le long de cette courbe, le champ de direction garde la pente constante p : on peut garnir l'isocline de petits segments de pente p. En traçant quelques isoclines pour différentes valeurs de p, on a un aperçu du champ de direction et de l'allure des solutions .

L'isocline pour la pente p est la courbe d'équation $f(t,x) = p$.

Exemple 1. Considérons l'équation $x' = x - t$. L'isocline pour la pente p est la droite d'équation $x = t + p$, parallèle à la première bissectrice des axes. La figure ci-contre montre les isoclines pour les pentes $p = 2,\ 1,\ 0$ et -1.



► L'isocline pour la pente 0 est la bissectrice d'équation $x = t$.

Dans le demi-plan $x > t$ (c'est-à-dire au dessus de la bissectrice), les solutions sont croissantes ; en dessous, elles sont décroissantes.

► Si une solution passe par un point de la bissectrice, elle y a un maximum.

► L'isocline pour la pente 1 est la droite $x = t+1$ de pente 1 : par suite, la fonction $u(t) = t+1$ est solution.

Exemple 2 : l'équation logistique. Voici un modèle très simple pour l'évolution d'une population en milieu fermé (par exemple, des bactéries en culture ou une population de lapins sur un îlot). Les observations effectuées par unité de temps δt (la durée d'un cycle reproductif) conduisent aux hypothèses suivantes :

► l'augmentation de population due à la reproduction est proportionnelle à l'effectif initial ;

► la compétition pour la nourriture induit une mortalité proportionnelle au carré de l'effectif initial.

Appelons $x(t)$ l'effectif à l'instant t.

On a ainsi $x(t+\delta t) - x(t) = kx(t)\delta t - a\big(x(t)\big)^2\delta t$, où k et a sont des constantes positives. Pour avoir un modèle continu, on fait tendre δt vers 0 : le rapport $\dfrac{x(t+\delta t) - x(t)}{\delta t}$ tend vers la dérivée $x'(t)$, vitesse d'évolution de la population, d'où la relation $x'(t) = kx(t) - a\big(x(t)\big)^2$.

Ainsi, la fonction $x(t)$ est solution de l'équation différentielle

$$x' = kx - ax^2 \qquad (\textit{équation logistique})$$

Supposons $k = 1{,}8$ et $a = 0{,}04$. L'équation s'écrit $x' = 1{,}8\,x - 0{,}04\,x^2$, ou encore

$$x' = 0{,}04\,x(45 - x)$$

L'isocline pour la pente 0 a pour équation $0{,}04\,x(45-x) = 0$, dont les solutions sont $x = 0$ et $x = 45$: l'isocline pour la pente 0 est donc formée de deux droites, l'axe des abscisses ($x = 0$) et la droite horizontale d'équation $x = 45$. Puisque le champ de directions est horizontal sur chacune de ces droites horizontales, on obtient deux solutions constantes, correspondant à des effectifs stables au cours du temps :

- la fonction $x(t) = 0$ quel que soit t,
- et la fonction $x(t) = 45$ quel que soit t ; si à un instant initial, la population est de 45 individus, l'effectif reste stable.

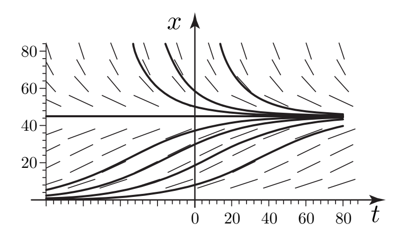
En tout point tel que $x{=}m$, le champ de directions a pour pente $p(m){=}0{,}04m(45{-}m)$:

la droite d'équation $x = m$ est l'isocline pour la pente $p(m)$.

Si $0<m<45$, alors $p(m)$ est positif : les solutions sont donc croissantes dans le domaine $0<x<45$. Si $m>45$, on a $p(m)<0$: les solutions sont décroissantes dans le domaine

$x > 45$. Voici des valeurs de $p(m)$, le champ de directions et quelques solutions :

m	5	10	20	30	40	45	50	60
$p(m)$	8	14	20	18	8	0	−10	−36



On observe que, mis à part la solution nulle, toutes les solutions $x(t)$ ont pour limite 45 quand t tend vers $+\infty$: quel que soit l'effectif initial, pourvu qu'il soit non nul, la population tend vers la valeur stable de 45 individus. Nous résoudrons analytiquement l'équation page 448.

1.2 Propriétés générales des solutions

Définition

Une *solution* de l'équation différentielle $x' = f(t, x)$ est une fonction $x(t)$ dérivable et définie sur un intervalle ouvert J, telle que $x'(t) = f\big(t, x(t)\big)$ pour tout $t \in J$.

Pour déterminer une solution $x(t)$, il suffit le plus souvent de choisir un nombre t_0 et la valeur x_0 de la solution pour $t = t_0$: on dit que (t_0, x_0) est une *condition initiale*. Notons D le *domaine de l'équation* $x' = f(t, x)$, c'est-à-dire le domaine de définition de la fonction f, et supposons f continue sur D.

Théorème

a) *Si deux solutions passent par un même point, leurs graphes sont tangents en ce point.*

b) *Supposons que la fonction f a des dérivées partielles $\dfrac{\partial f}{\partial x}$ et $\dfrac{\partial f}{\partial t}$ continues. Pour tout point (t_0, x_0) de D, il existe une unique solution $x(t)$ telle que $x(t_0) = x_0$.*

Une solution est donc entièrement déterminée par sa condition initiale.

c) *Le graphe de chaque solution va jusqu'au bord du domaine D (éventuellement à l'infini).*

La propriété (a) est évidente : si $A = (\tau, a)$ est un point commun aux graphes de deux solutions x_1 et x_2, la tangente en A au graphe de x_1 a pour pente $f\big(\tau, a\big)$ et il en va de même de la tangente en A au graphe de x_2 : ces deux graphes sont donc tangents au point A.

L'énoncé (b) affirme deux propriétés :

- Il existe toujours une solution $x(t)$ satisfaisant la condition initiale $x(t_0) = x_0$,
- et cette solution est unique : cela veut dire que si $x_1(t)$ et $x_2(t)$ sont des solutions prenant la même valeur x_0 en $t = t_0$, alors on a $x_1(t) = x_2(t)$ quel que soit t.

Conséquence. *Si f a des dérivées partielles continues dans D, alors les graphes de deux solutions différentes n'ont aucun point commun.*

Dans ce cas, les graphes des solutions forment une partition du domaine de l'équation : c'est ce qu'on voit sur les figures précédentes. À chaque fois que nous pourrons calculer explicitement les solutions d'une équation différentielle, nous vérifierons toutes ces propriétés.

Définition

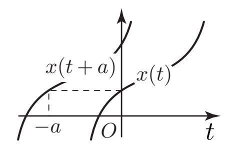
Une équation différentielle de la forme $x' = f(x)$ est dite *autonome*.

Si l'on imagine que t représente le temps, une équation différentielle autonome exprime que la relation entre la fonction x et sa dérivée ne dépend pas du temps. Dans les applications, on rencontre de nombreuses équations différentielles autonomes. L'équation logistique étudiée dans le paragraphe précédent est autonome.

Proposition. *Si $x(t)$ est solution d'une équation différentielle autonome, alors pour tout nombre a, la fonction $y(t) = x(t+a)$ est solution.*

Démonstration. On a en effet $y'(t) = x'(t+a) = f\big(x(t+a)\big) = f\big(y(t)\big)$, donc la fonction y est solution de l'équation différentielle. ■

Traduction géométrique : Si $x(t)$ est une fonction et a un nombre, le graphe de la fonction $t \mapsto x(t+a)$ s'obtient en translatant le graphe de $x(t)$ de la quantité $-a$ parallèlement à l'axe des t. La propriété ci-dessus se formule donc de la manière suivante.



Si l'équation est autonome, alors en translatant parallèlement à l'axe des t le graphe d'une solution, on obtient encore le graphe d'une solution.

1.3 Équations différentielles linéaires

Définitions

Une *équation différentielle linéaire du premier ordre* est de la forme $x' = a(t)x + b(t)$, où $a(t)$ et $b(t)$ sont des fonctions continues sur un même intervalle I. L'équation différentielle $x' = a(t)x$ s'appelle *l'équation homogène associée* et la fonction b s'appelle le *second membre*.

Le domaine de l'équation est l'ensemble des points (t, x) tels que $t \in I$.

Résolution de l'équation homogène $x' = a(t)x$

Soit $A(t)$ une primitive de $a(t)$ sur l'intervalle I.

Proposition. *Les solutions de l'équation homogène $x' = a(t)x$ sont les fonctions $u(t) = K \exp\big[A(t)\big]$, où K est un nombre quelconque.*

Démonstration. Soit $u(t)$ une fonction dérivable sur I. Posons $y(t)=u(t)\exp\big[-A(t)\big]$, de sorte que $u(t)=y(t)\exp\big[A(t)\big]$. Puisque $\exp\big[A(t)\big]$ a pour dérivée $A'(t)\exp\big[A(t)\big]=a(t)\exp\big[A(t)\big]$, on a

$$u'(t) = y'(t)\exp\big[A(t)\big] + a(t)y(t)\exp\big[A(t)\big] = y'(t)\exp\big[A(t)\big] + a(t)u(t)\,.$$

La fonction u est donc solution de $x' = a(t)x$ si et seulement si $y'(t) = 0$ pour tout $t \in I$, c'est-à-dire si et seulement si y est constante sur I. ■

Remarque

Cette résolution s'applique à l'équation $z' = az$, où a est un nombre complexe : puisque la dérivée de $z(t) = e^{at}$ est $z'(t) = ae^{at} = az(t)$ (proposition page 321), la démonstration ci-dessus montre que les solutions à valeurs complexes de l'équation $z' = az$ sont les fonctions $t \mapsto ke^{at}$, où k est un nombre complexe quelconque.

Résolution de l'équation complète $x' = a(t)x + b(t)$

Proposition. *Si $s(t)$ est une solution de l'équation $x' = a(t)x + b(t)$, alors les solutions de cette équation sont les fonctions $x(t) = K\exp\big[A(t)\big] + s(t)$, où $A(t)$ est une primitive de la fonction a et K une constante quelconque.*

Les solutions de $x'=a(t)x+b(t)$ s'obtiennent en ajoutant à une solution particulière $s(t)$, une solution quelconque de l'équation homogène $x' = a(t)x$.

Démonstration. Soit $x(t)$ une fonction dérivable. Posons $y(t) = x(t) - s(t)$, donc $x(t) = y(t) + s(t)$. On a $s'(t) - a(t)s(t) = b(t)$, donc

$$y'(t) - a(t)y(t) = x'(t) - s'(t) - a(t)x(t) + a(t)s(t) = x'(t) - a(t)x(t) - b(t)$$

Pour que $x(t)$ soit solution de l'équation $x'=a(t)x+b(t)$, il faut et il suffit que $y'(t)-a(t)y(t)=0$ quel que soit $t \in I$, c'est-à-dire que y soit solution de l'équation homogène. ■

D'après la proposition, il suffit de trouver une solution particulière $s(t)$ de l'équation pour en déduire toutes les solutions. Par exemple, pour l'équation $x' = x-t$ de l'exemple 1 page 440, les solutions sont toutes les fonctions $x(t) = Ke^t + t + 1$, car $s(t) = t+1$ est une solution particulière.

Recherche d'une solution particulière. Supposons qu'on a calculé une primitive $A(t)$ de la fonction $a(t)$, donc aussi une solution non nulle $u(t)$ de l'équation homogène. Cherchons une solution $s(t)$ de $x' = a(t)x + b(t)$ sous la forme

$$s(t) = \lambda(t)u(t)\ , \quad \text{où } \lambda(t) \text{ est une fonction à déterminer.}$$

On a $u'(t) = a(t)u(t)$ donc

$$s'(t) = \lambda(t)u'(t) + \lambda'(t)u(t) = \lambda(t)a(t)u(t) + \lambda'(t)u(t)$$

$$a(t)s(t) + b(t) = a(t)\lambda(t)u(t) + b(t)$$

Pour que $s'(t) = a(t)s(t) + b(t)$, il faut et il suffit que $\lambda'(t)u(t) = b(t)$. Puisque

$$1/u(t) = \exp\big[-A(t)\big]\,,$$

cette condition s'écrit

$$\lambda'(t) = b(t)\exp\big[-A(t)\big]$$

Il suffit ainsi de calculer une primitive $\lambda(t) = \int^t b(\tau) \exp\big[-A(\tau)\big]\, d\tau$ et la fonction $\lambda(t)u(t)$ sera une solution particulière de l'équation $x' = a(t)x + b(t)$.
Cette méthode de recherche d'une solution particulière s'appelle la *méthode de variation de la constante*.

Détermination d'une solution au moyen des conditions initiales

Donnons-nous des conditions initiales $t_0 \in I$ et x_0 quelconque.

Première méthode. Les solutions s'écrivent $x(t) = K \exp\big[A(t)\big] + s(t)$, où $s(t)$ est une solution particulière et K un nombre quelconque. Pour que $x(t_0) = x_0$, il suffit donc de choisir le nombre K tel que $x_0 = K \exp\big[A(t_0)\big] + s(t_0)$.

Seconde méthode. Choisissons pour A la primitive de $a(t)$ telle que $A(t_0) = 0$.

- Si l'on pose $\lambda(t) = \int_{t_0}^t b(\tau) \exp\big[-A(\tau)\big]\, d\tau$, alors $\lambda(t_0) = 0$ et la solution particulière $s(t) = \lambda(t) \exp\big[A(t)\big]$ vérifie $s(t_0) = 0$.
- On sait que les solutions sont les fonctions $x(t) = K \exp\big[A(t)\big] + s(t)$, où K est une constante. Puisque $A(t_0) = s(t_0) = 0$, il vient $x(t_0) = K$.

Puisque $A(t) = \int_{t_0}^t a(\tau)\, d\tau$, la solution telle que $x(t_0) = x_0$ est donc

$$x(t) = x_0 \exp\big[A(t)\big] + \lambda(t) \exp\big[A(t)\big] \ , \quad \text{où } \lambda(t) = \int_{t_0}^t b(\tau) \exp\big[-A(\tau)\big]\, d\tau$$

Exemple 1. Le refroidissement d'une quantité de matière dans un courant d'air est proportionnel à la différence de température avec l'air. Soit θ la température de l'air, maintenue constante. Supposons que le corps est à la température initiale T_0. Si l'on note $T(t)$ sa température au bout d'un temps t, on a donc

$$T'(t) = -a\big[T(t) - \theta\big] \ , \quad \text{où } a \text{ est un coefficient positif.}$$

En posant $x(t) = T(t) - \theta$, il vient $x'(t) = T'(t) = -a\big[T(t) - \theta\big] = -ax(t)$, donc $x(t) = \lambda e^{-at}$, où λ est une constante. Puisque $T(t) = x(t) + \theta$, on obtient une solution de la forme $T(t) = \lambda e^{-at} + \theta$.

Supposons que l'air est à 20 degrés et que le corps passe de 100 degrés à 60 degrés en un quart d'heure. En comptant le temps en heures, on a donc $T(0) = 100$ et $T(1/4) = 60$.
Il vient $100 = T(0) = \lambda + 20$, d'où $\lambda = 80$, puis $60 = T(1/4) = \lambda e^{-a/4} + 20 = 80 e^{-a/4} + 20$, d'où $e^{-a/4} = 1/2$ et $a = 4 \ln 2$.
Finalement, la loi de refroidissement du corps est

$$T(t) = 80 e^{-(4 \ln 2)t} + 20, \quad \text{où } t \text{ est compté en heures}$$

Exemple 2. Considérons un circuit électrique formé d'une résistance R, d'un condensateur de capacité C et d'un générateur de force électromotrice $E(t) = E_m \sin \omega t$ placés en série. Si le condensateur est initialement déchargé, sa charge $q(t)$ à l'instant

t est solution de l'équation différentielle

$$R\frac{dq}{dt} + \frac{q}{C} = E_m \sin \omega t$$

Résolution de l'équation homogène $q' = \frac{-1}{RC} q$. Les solutions sont les fonctions $K \exp\left(\frac{-t}{RC}\right)$, où K est une constante quelconque.

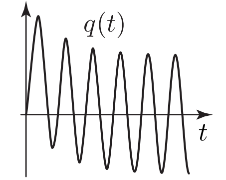
Recherche d'une solution particulière. Elle est de la forme $s(t) = \lambda(t) \exp\left(\frac{-t}{RC}\right)$, où $\lambda'(t) \exp\left(\frac{-t}{RC}\right) = \frac{E_m}{R} \sin \omega t$. Il vient $\lambda'(t) = \frac{E_m}{R} \exp\left(\frac{t}{RC}\right) \sin \omega t$. Les primitives $\lambda(t)$ sont de la forme (page 321)

$$\big(a \sin \omega t + b \cos \omega t\big) \exp\left(\frac{t}{RC}\right)$$

En identifiant, on obtient $a = \frac{E_m C}{1 + (RC\omega)^2}$ et $b = -\frac{E_m C(RC\omega)}{1 + (RC\omega)^2}$ et la fonction $s(t) = a \sin \omega t + b \cos \omega$ est donc une solution particulière.

La solution générale de l'équation est ainsi $q(t) = K \exp\left(\frac{-t}{RC}\right) + a \sin \omega t + b \cos \omega t$. La condition initiale $q(0) = 0$ s'écrit $0 = K + b$, d'où la fonction de charge du condensateur :

$$q(t) = \frac{E_m C}{1 + (RC\omega)^2} \left[RC\omega \exp\left(\frac{-t}{RC}\right) + \sin \omega t - RC\omega \cos \omega t\right]$$



Comme les valeurs de $\exp(-t/RC)$ deviennent en général rapidement négligeables, on voit s'installer un régime permanent quasi-périodique, de quasi-période celle du générateur.

1.4 Équations différentielles à variables séparées

Définition

Si f et g sont des fonctions, une équation différentielle de la forme $x' = f(x)g(t)$ est dite *à variables séparées*.

Une équation autonome est à variables séparées.

Exemple. Résolvons l'équation différentielle $x' = -2tx^2$ dont nous avons dessiné le champ de directions page 440. La fonction nulle : $x(t) = 0$ quel que soit t, est une solution. Cherchons maintenant les solutions dans le domaine $x \neq 0$.

L'équation s'y écrit $\frac{x'(t)}{x(t)^2} = -2t$. Puisqu'on a $\frac{-x'(t)}{x(t)^2} = \frac{d}{dt}\left(\frac{1}{x(t)}\right)$, il vient $\frac{d}{dt}\left(\frac{1}{x(t)}\right) = 2t = \frac{d}{dt}(t^2)$, donc $\frac{1}{x(t)} = t^2 + C$, où C est une constante. Ainsi, dans

le domaine $x \neq 0$, les solutions sont

$$x(t) = \frac{1}{t^2 + C}, \quad \text{avec } C \text{ constant.}$$

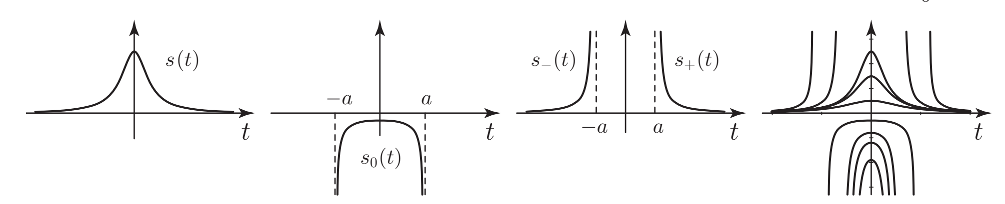
Nous avons déjà remarqué que le sens de variation des solutions d'une équation différentielle est donné par le signe de $f(t,x)$: ici, les solutions sont croissantes pour $t<0$ et décroissantes pour $t>0$. L'intervalle de définition d'une solution dépend du signe de C.

Supposons $C > 0$. La solution $s(t) = \dfrac{1}{t^2 + C}$ est positive, définie sur $\mathbb{R}$ et tend vers 0 quand t tend vers $\pm\infty$; le maximum est atteint en $t = 0$ (figures ci-dessous).

Supposons $C < 0$. Posons $C = -a^2$, avec $a > 0$. Rappelons que, par définition, une solution est toujours définie sur un intervalle. La formule de résolution détermine trois intervalles de définition : $I_1 =]-a, a[$, $I_2 =]-\infty, -a[$ et $I_3 =]a, +\infty[$, donc trois types de solutions.

- Posons $s_0(t) = \dfrac{1}{t^2 - a^2}$ pour $t \in]-a, a[$. Cette solution s_0 est à valeurs négatives et possède des asymptotes verticales en $t = \pm a$: on a $\lim\limits_{t \to a} s_0(t) = \lim\limits_{t \to -a} s_0(t) = -\infty$.
- Posons $s_+(t) = \dfrac{1}{t^2 - a^2}$ pour $t \in]a, +\infty[$. C'est une solution à valeurs positives, décroissante et de limite 0 quand t tend vers $+\infty$; puisque $s_+(t)$ tend vers $+\infty$ quand t tend vers a par valeurs supérieures, cette solution a une asymptote verticale en a.
- De même, la fonction $s_-(t) = \dfrac{1}{t^2 - a^2}$ pour $t \in]-\infty, -a[$, est une solution à valeurs positives, croissante de 0 à $+\infty$.

La seule solution qui satisfait une condition initiale $x(t_0) = 0$ est la fonction nulle. Si $x_0 \neq 0$, la solution telle que $x(t_0) = x_0$ s'obtient en calculant C pour que $x_0 = \dfrac{1}{t_0^2 + C}$.



En observant simplement le champ de directions, il n'est pas commode de deviner les solutions s_+ et s_-.

Résolution d'une équation $x' = f(x)g(t)$

Supposons que, dans l'équation, les fonctions f et g ont des dérivées continues.

a) On commence par chercher les solutions constantes. La dérivée d'une telle solution $x(t)$ est nulle, donc en appelant m sa valeur, il vient $0 = x'(t) = f\big(x(t)\big)g(t) = f(m)g(t)$ pour tout t. Si g n'est pas la fonction nulle, on a donc $f(m) = 0$.

Toute solution constante a pour valeur un nombre m tel que $f(m) = 0$.

b) Les solutions non constantes vérifient $\dfrac{dx}{dt}\dfrac{1}{f\big(x(t)\big)} = g(t)$ ou encore

$$\frac{dx}{f(x)} = g(t)\,dt$$

La solution $s(t)$ telle $s(t_0) = x_0$ s'obtient en intégrant chaque membre :

$$\int_{x_0}^{s(t)} \frac{dx}{f(x)} = \int_{t_0}^{t} g(\tau)\,d\tau\,,$$

ce qui exprime la relation entre $s(t)$ et t.

Exemple : l'équation logistique. Il s'agit de l'équation autonome (page 441)

$$x' = kx - ax^2\ , \text{ où } k \text{ et } a \text{ sont des nombres positifs.}$$

Menons la résolution dans le domaine $x > 0$.

Recherche des solutions constantes. La valeur d'une solution constante est un nombre $m > 0$ vérifiant $km - am^2 = 0$, donc $m = k/a$: il y a une seule solution constante, définie par $x(t) = k/a$.

Recherche des autres solutions. Puisque les graphes de deux solutions différentes n'ont aucun point commun, les solutions non constantes ne prennent pas la valeur k/a : leur graphe est donc contenu dans l'un des domaines $0<x<k/a$ ou $x>k/a$. Soit $s(t)$ la solution telle que $s(0) = x_0$, où x_0 est un nombre positif différent de k/a. Si l'on a $0 < x_0 < k/a$, alors le graphe de s est dans le domaine $0 < x < k/a$; si $x_0 > k/a$, le graphe de s est dans le domaine $x > k/a$.
En écrivant l'équation sous la forme $\dfrac{dx}{kx-ax^2} = dt$, la fonction $s(t)$ est déterminée par la relation

(1) $$\int_{x_0}^{s(t)} \frac{dx}{kx - ax^2} = \int_0^t d\tau = t$$

Pour calculer l'intégrale, cherchons les nombres α et β tels que $\dfrac{1}{x(k-ax)} = \dfrac{\alpha}{x} + \dfrac{\beta}{k-ax}\cdot$
En identifiant, on trouve $\alpha = 1/k$, $\beta = a/k$ et

$$\frac{1}{x(k-ax)} = \frac{1}{k}\left[\frac{1}{x} + \frac{a}{k-ax}\right]$$

Il vient $\displaystyle\int \frac{dx}{kx - ax^2} = \frac{1}{k}\big[\ln x - \ln|k-ax|\big] = \frac{1}{k}\ln\frac{x}{|k-ax|}$ et (1) s'écrit

$$\frac{1}{k}\left[\ln\frac{s(t)}{|k-as(t)|} - \ln\frac{x_0}{|k-ax_0|}\right] = \frac{1}{k}\ln\left[\frac{s(t)}{|k-as(t)|}\left(\frac{x_0}{|k-ax_0|}\right)^{-1}\right] = t$$

En posant $C = \dfrac{x_0}{|k-ax_0|} > 0$, on obtient

(2) $$\frac{s(t)}{|k-as(t)|} = Ce^{kt}$$

Supposons $0 < x_0 < k/a$. On a alors $0 < s(t) < k/a$ pour tout t, c'est-à-dire $k - as(t) > 0$. Il vient $\frac{s(t)}{k - as(t)} = Ce^{kt}$, d'où

$$s(t) = \frac{Ck}{Ca + e^{-kt}}\ , \quad \text{pour tout } t \in \mathbb{R}.$$

Quand t va de $-\infty$ à $+\infty$, e^{-kt} décroît de $+\infty$ à 0, donc $s(t)$ croît de 0 jusqu'à $k/a = \lim_{t \to +\infty} s(t)$.

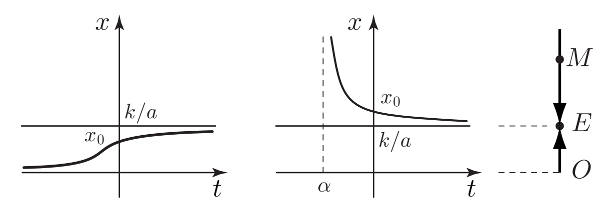
Supposons $x_0 > k/a$. On a $s(t) > k/a$ pour tout t, donc $k - as(t) < 0$ et $|k - as(t)| = as(t) - k$. On obtient ainsi

$$s(t) = \frac{Ck}{Ca - e^{-kt}}\ , \quad \text{pour } t > -\frac{1}{k}\ln(Ca).$$

Puisque $Ca = \frac{ax_0}{ax_0 - k}$ et $ax_0 - k > 0$, on a $Ca > 1$ et la borne $\alpha = -\frac{1}{k}\ln(Ca)$ de l'intervalle de définition est un nombre négatif : la solution est donc bien définie en $t = 0$. Quand t tend vers α (par valeurs supérieures), le dénominateur tend vers 0 et $s(t)$ tend vers $+\infty$: la solution a une asymptote verticale en α. Quand t tend vers $+\infty$, e^{-kt} tend vers 0 et $s(t)$ tend vers k/a.

Dans le domaine $x > 0$, toutes les solutions ont pour limite k/a quand t tend vers $+\infty$: la population tend vers l'effectif stable de k/a individus, comme nous l'avions observé sur le champ de directions page 442.

Puisque l'équation logistique est autonome, si l'on translate le graphe d'une solution parallèlement à l'axe du temps, on obtient encore le graphe d'une solution.



Sur la figure de droite, l'effectif $x(t)$ est représenté par un point M sur une demi-droite d'origine O. Si l'on se donne la position initiale, le point M décrit une trajectoire sur cette demi-droite. On voit qu'il y a trois trajectoires possibles :

- Si la position initiale est le point d'équilibre E d'abscisse k/a, alors M reste en E : l'ensemble $\{E\}$ est une trajectoire.
- L'intervalle $T_- =]O; E[$ est la trajectoire correspondant à une position initiale située entre E et l'origine. Un point qui décrit cette trajectoire tend vers l'équilibre E quand le temps tend vers l'infini.
- Pour une position initiale située au delà de E, la trajectoire est la demi-droite ouverte limitée par E. Sur cette trajectoire, le point tend aussi vers l'équilibre quand le temps tend vers l'infini.

2. Équations différentielles linéaires d'ordre 2

Nous allons étudier les équations différentielles linéaires faisant intervenir la fonction x ainsi que ses deux premières dérivées x' et x''. Ce sont des équations très courantes.

2.1 Propriétés générales

Définitions

Une *équation différentielle linéaire d'ordre* 2 est de la forme $x'' + p(t)x' + q(t)x = b(t)$, où $p(t)$, $q(t)$ et $b(t)$ sont des fonctions continues sur un même intervalle I. La fonction b s'appelle le *second membre* de l'équation. L'équation $x''+p(t)x'+q(t)x=0$ est l'équation *homogène* associée.

Propriétés des solutions

1) Soit $x(t)$ une solution de l'équation homogène (h) $x''+p(t)x'+q(t)x = 0$. Posons $y(t) = \lambda x(t)$, où λ est un nombre. On a

$$y'' + p(t)y' + q(t)y = \lambda x'' + \lambda p(t)x' + \lambda x = \lambda\big(x'' + p(t)x' + q(t)x\big) = 0\,,$$

donc la fonction $y(t)$ est solution de (h).

2) Supposons que $x_1(t)$ et $x_2(t)$ sont des solutions de (h). Alors la fonction $y(t) = x_1(t) + x_2(t)$ est solution de (h), car

$$y''+p(t)y'+q(t)y = \big(x_1''+p(t)x_1'+q(t)x_1\big) + \big(x_2''+p(t)x_2'+q(t)x_2\big) = 0\,.$$

3) Supposons que $s(t)$ est une solution de l'équation (e) $x''+p(t)x'+q(t)x = b(t)$. Pour qu'une fonction $x(t)$ soit aussi une solution de (e), il faut et il suffit que l'on ait $x'' + p(t)x' + q(t)x = s'' + p(t)s' + q(t)s$, c'est-à-dire

$$(x-s)'' + p(t)(x-s)' + q(t)(x-s) = 0\,,$$

ce qui veut dire que $u = x-s$ est solution de (h). Les solutions de (e) sont donc les fonctions $x(t) = u(t) + s(t)$, où $u(t)$ est solution de (h).

Théorème

a) *Soient $x_1(t)$ et $x_2(t)$ des solutions de l'équation homogène (h) $x''+p(t)x'+q(t)x = 0$.*
   - *Pour tous nombres λ_1, λ_2, la fonction $\lambda_1 x_1(t) + \lambda_2 x_2(t)$ est solution de (h).*
   - *Si les fonctions x_1 et x_2 ne sont pas proportionnelles, alors les solutions de (h) sont les fonctions $u(t) = \lambda_1 x_1(t) + \lambda_2 x_2(t)$, où λ_1 et λ_2 sont des nombres quelconques.*

b) *Supposons que $s(t)$ est une solution de l'équation complète (e) $x''+p(t)x'+q(t)x=b(t)$. Alors les solutions de (e) sont les fonctions $u(t) + s(t)$, où $u(t)$ est une solution quelconque de l'équation homogène (h).*

c) *Soient $t_0 \in I$ et x_0, v_0 des nombres quelconques. Il y a une unique solution $s(t)$ de (e) satisfaisant les conditions initiales $s(t_0) = x_0$, $s'(t_0) = v_0$.*

Pour résoudre l'équation (e), on cherche les solutions $\lambda_1 x_1(t) + \lambda_2 x_2(t)$ de l'équation homogène et une solution particulière de (e). On calcule ensuite les constantes λ_1 et λ_2 pour satisfaire les conditions initiales données.

Il n'y a pas de méthode générale pour résoudre l'équation, mais on trouvera quelques techniques d'étude dans les exercices en fin de chapitre.

Calcul d'une solution particulière. Supposons qu'on ait trouvé des solutions non proportionnelles $u(t)$ et $v(t)$ de l'équation homogène (h). Voici comment obtenir une solution particulière s de l'équation (e).

i) On calcule la fonction $w = uv' - u'v$ (elle ne s'annule pas),

ii) et les fonctions $\lambda(t) = \int_{t_0}^{t} \frac{-b(s)v(s)}{w(s)}\, ds$ et $\mu(t) = \int_{t_0}^{t} \frac{b(s)u(s)}{w(s)}\, ds$.

Alors $s(t)=\lambda(t)u(t)+\mu(t)v(t)$ est une solution de (e) telle que $s(t_0)=0$ (voir page 493).

Aspect vectoriel

Si x_1 et x_2 sont solutions de (h), alors d'après les propriétés (1) et (2), il en va de même de la combinaison linéaire $\lambda_1 x_1 + \lambda_2 x_2$.

L'ensemble des solutions d'une équation linéaire homogène est un espace vectoriel.

Pour toute fonction $x(t)$ deux fois dérivable sur I, notons $L(x)$ la fonction

$$t \mapsto x''(t) + p(t)x'(t) + q(t)x(t)\,.$$

On définit ainsi une application $L : V \to F$, où V est l'espace vectoriel des fonctions deux fois dérivables sur l'intervalle I et F l'espace de toutes les fonctions. On dit que L est un *opérateur différentiel*. On a $L(\lambda x) = \lambda L(x)$ pour tout $x \in V$ et $L(x_1 + x_2) = L(x_1) + L(x_2)$ pour tous $x_1 \in V$, $x_2 \in V$.

L'opérateur L est une application linéaire.

Par définition, une solution de (h) est une fonction $x \in V$ telle que $L(x) = 0$: l'ensemble des solutions de (h) est donc le noyau de L (page 172).
Traduisons la seconde affirmation (a) du théorème en notant $S_0 = \operatorname{Ker} L$ l'espace vectoriel des solutions de (h) : si x_1 et x_2 sont des vecteurs de S_0 non colinéaires, alors tout vecteur $x \in S_0$ est combinaison linéaire de x_1 et x_2 ; cela veut dire que (x_1, x_2) est une base de S_0.

L'espace vectoriel des solutions de (h) est de dimension 2.

Les solutions de (e) sont les fonctions $x \in V$ telles que $L(x) = b$. Supposons que s est une solution de l'équation (e). Pour toute fonction $x \in V$, on a les équivalences

$$L(x) = b \iff L(x) = L(s) \iff L(x - s) = 0$$

Cela exprime que les solutions de (e) sont les fonctions de la forme $u + s$, où $u \in S_0$: c'est l'affirmation (b) du théorème.
Soit $t_0 \in I$. Notons S l'ensemble des solutions de (e) et définissons l'application $C : S \to \mathbb{R}^2$ qui à toute solution $x(t)$ de (e) associe le couple $C(x) = \big(x(t_0), x'(t_0)\big)$.

Pour tout $(x_0, v_0) \in \mathbb{R}^2$, il y a une et une seule solution $x(t)$ ayant pour conditions initiales $x(t_0) = x_0$, $x'(t_0) = v_0$. Pour l'application C, la fonction $x(t)$ est donc l'unique antécédent de (x_0, v_0). Cela montre que l'application C est une bijection de S sur $\mathbb{R}^2$.

Principe de superposition. Supposons que le second membre $b(t)$ est une somme $b_1(t) + b_2(t)$ de deux fonctions.

Si $s_1(t)$ est une solution de l'équation $x''+p(t)x'+q(t)x = b_1(t)$ et si $s_2(t)$ est une solution de l'équation $x''+p(t)x'+q(t)x = b_2(t)$, alors $s_1(t) + s_2(t)$ est solution de l'équation $x''+p(t)x'+q(t)x = b_1(t) + b_2(t)$.

2.2 Équations linéaires à coefficients constants

Il s'agit des équations différentielles de la forme

$$(e) \quad x'' + px' + qx = b(t)$$

où p et q sont des nombres réels et $b(t)$ une fonction continue sur un intervalle.

Pour résoudre l'équation, il suffit, d'après les propriétés générales, de trouver une solution particulière $s(t)$ et de résoudre l'équation homogène $x'' + px' + qx = 0$.

Résolution de l'équation homogène (h) $x'' + px' + qx = 0$

Il est plus simple de chercher les solutions à valeurs complexes. Rappelons que pour dériver une fonction à valeurs complexes, on dérive sa partie réelle et sa partie imaginaire. Une telle fonction est donc solution si et seulement si sa partie réelle et sa partie imaginaire le sont (page 321).

Rappelons aussi que si $z = a + b\boldsymbol{i}$ est un nombre complexe, on a posé $e^{zt} = e^{(a+b\boldsymbol{i})t} = e^{at}\big(\cos bt + \boldsymbol{i}\sin bt\big)$, pour tout nombre réel t. La fonction $t \mapsto e^{zt}$ a pour dérivée ze^{zt}.

Posons $x(t) = e^{zt}$. Puisque $x''(t) = z^2 e^{zt}$, il vient $x''(t) + px'(t) + qx(t) = (z^2 + pz + q)e^{zt}$.

La fonction e^{zt} est solution de (h) si et seulement si le nombre z satisfait l'égalité $z^2 + pz + q = 0$.

Introduisons *l'équation caractéristique* $z^2 + pz + q = 0$.

Premier cas : $p^2 - 4q > 0$. L'équation caractéristique a deux racines réelles distinctes r_1 et r_2, d'où les solutions $u_1(t) = e^{r_1 t}$ et $u_2(t) = e^{r_2 t}$. Ces fonctions n'étant pas proportionnelles, on en déduit d'après le théorème précédent :

les solutions de (h) sont les fonctions $\lambda_1 e^{r_1 t} + \lambda_2 e^{r_2 t}$, où λ_1 et λ_2 sont des nombres réels quelconques.

Deuxième cas : $p^2-4q<0$. L'équation caractéristique a deux racines distinctes conjuguées $r+\boldsymbol{i}\,\omega$ et $r-\boldsymbol{i}\,\omega$.
Les fonctions $x(t)=e^{rt}e^{\boldsymbol{i}\,\omega t}$ et $\bar{x}(t)=e^{rt}e^{-\boldsymbol{i}\,\omega t}$ sont donc des solutions de (h) et il en va de même des fonctions

$$u_1(t)=(1/2)\big(x(t)+\bar{x}(t)\big)=e^{rt}\cos\omega t \quad \text{et} \quad u_2(t)=(1/2\,\boldsymbol{i})\big(x(t)-\bar{x}(t)\big)=e^{rt}\sin\omega t.$$

Puisque u_1 et u_2 ne sont pas proportionnelles, on en déduit :

les solutions de (h) sont les fonctions $e^{rt}\big(\lambda_1\cos\omega t+\lambda_2\sin\omega t\big)$,
où λ_1 et λ_2 sont des nombres réels quelconques.

Troisième cas : $p^2-4q=0$. L'équation caractéristique a une racine double $-p/2$, ce qui fournit la solution réelle $u_1(t)=e^{-(p/2)t}$. Montrons que la fonction $u_2(t)=tu_1(t)$ est aussi une solution : on a en effet $u_2'(t)=tu_1'(t)+u_1(t)$, $u_2''(t)=tu_1''(t)+2u_1'(t)$ et

$$\begin{aligned} u_2''(t)+pu_2'(t)+qu_2(t) &= tu_1''(t)+2u_1'(t)+ptu_1'(t)+pu_1(t)+qtu_1(t) \\ &= t\big[u_1''(t)+pu_1'(t)+qu_1(t)\big]+2u_1'(t)+pu_1(t)=0 \end{aligned}$$

car u_1 est solution et $u_1'=-(p/2)u_1$. Puisque les solutions u_1 et u_2 ne sont pas proportionnelles,

les solutions de (h) sont les fonctions $(\lambda_1+\lambda_2 t)e^{-(p/2)t}$, où
λ_1 et λ_2 sont des nombres réels quelconques.

Comme nous n'avons pas démontré le théorème général du précédent paragraphe, vérifions que l'on obtient bien ainsi toutes les solutions.
Dans l'équation (h), faisons le changement de fonction inconnue $x(t)=e^{rt}y(t)$, où $r=-p/2$. Puisque $x'(t)=\big(y'(t)+ry(t)\big)e^{rt}$ et $x''(t)=\big(y''(t)+2ry'(t)+r^2y(t)\big)e^{rt}$, il vient

$$x''(t)+px'(t)+qx(t)=\big[y''(t)+(2r+p)y'(t)+\big(r^2+pr+q\big)y(t)\big]e^{rt}$$

On a $2r+p=0$ et $r^2+pr+q=p^2/4-p^2/2+q=-(p^2-4q)/4$.
► Plaçons-nous dans le cas $p^2-4q\geqslant 0$ et posons $a=(1/2)\sqrt{p^2-4q}$. Il vient $r^2+pr+q=-a^2$ et $x''(t)+px'(t)+qx(t)=\big[y''(t)-a^2y(t)\big]e^{rt}$. Ainsi la fonction $x(t)$ est solution de (h) si et seulement si $y(t)$ est solution de l'équation différentielle $(*)$ $y''-a^2y=0$.
Les fonctions $y_1(t)=e^{at}$ et $y_2(t)=e^{-at}$ sont évidemment solutions de $(*)$. Soit $y(t)$ une solution quelconque.
Posons $z(t)=\big[y'(t)-ay(t)\big]e^{at}$. En dérivant, on obtient $z'(t)=\big(y''(t)-a^2y(t)\big)e^{at}=0$, donc on a $z(t)=c$, une constante. Ainsi $y'-ay=ce^{-at}$ et y est solution de l'équation différentielle $y'=ay+ce^{-at}$. Les solutions de l'équation homogène sont les fonctions λe^{at}.
Si $a\neq 0$, la fonction $-(c/2a)e^{-at}$ est une solution particulière, donc $y(t)=\lambda e^{at}-(c/2a)e^{-at}$.
Les solutions de (h) sont $x(t)=e^{rt}y(t)=\lambda^{(r+a)t}-(c/2a)e^{(r-a)t}$, où λ et c sont des constantes quelconques. Puisque les nombres $r\pm a$ sont les racines de l'équation caractéristique, on obtient bien le résultat énoncé.
Si $a=0$, c'est-à-dire $p^2-4q=0$, on a $y'=c$, $y(t)=ct+d$, d'où $x(t)=(ct+d)e^{rt}$, avec c et d des constantes.
► Dans le cas $p^2-4q<0$, on a $\omega=(1/2)\sqrt{4q-p^2}$, donc $r^2+pr+q=-(p^2-4q)/4=\omega^2$.
La fonction $x(t)$ est solution de (h) si et seulement si $y(t)$ est solution de l'équation différentielle $y''+\omega^2y=0$.

Posons $z(t)=\big(y'(t)-\boldsymbol{i}\omega y(t)\big)e^{\boldsymbol{i}\,\omega t}$. En dérivant, on obtient $z'(t)=\big(y''(t)+\omega^2 y(t)\big)e^{\boldsymbol{i}\,\omega t}=0$, donc $z(t)=c$, une constante. Comme ci-dessus, $y(t)$ est solution de l'équation différentielle du premier ordre (à coefficients complexes) $y'=\boldsymbol{i}\omega y+ce^{-\boldsymbol{i}\,\omega t}$. La fonction $-(c/2\boldsymbol{i}\omega)e^{-\boldsymbol{i}\,\omega t}$ est une solution particulière, d'où la solution générale : $y(t)=\alpha e^{\boldsymbol{i}\,\omega t}-(c/2\,\boldsymbol{i}\,\omega)e^{-\boldsymbol{i}\,\omega t}$, α et c étant des nombres complexes quelconques. Les solutions à valeurs réelles sont les $\lambda_1\cos\omega t+\lambda_2\sin\omega t$, où λ_1 et λ_2 sont réels, et en multipliant par e^{rt}, on obtient les solutions de (h).

Exemple : l'oscillateur libre

On nomme ainsi les dispositifs physiques conduisant à une équation différentielle linéaire homogène d'ordre 2. En voici deux exemples (on note comme d'habitude $\dot{x}$ et $\ddot{x}$ les dérivées par rapport au temps).

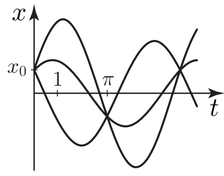
1) Considérons une masse m suspendue à un ressort de raideur k. Si l'on écarte la masse de sa position d'équilibre (dans le sens vertical), son mouvement est régi par l'équation différentielle $m\ddot{x}+c\dot{x}+kx=0$, où c est un coefficient d'amortissement (dû par exemple au frottement).

2) Quand on laisse se décharger un condensateur de capacité C dans une bobine d'inductance L et de résistance R, sa charge $q(t)$ satisfait l'équation différentielle $L\ddot{q}+R\dot{q}+\dfrac{1}{C}q=0$ (l'intensité dans le circuit est $\dot{q}(t)$).

Décrivons le comportement des solutions en prenant l'exemple du pendule d'équation (h) $m\ddot{x}+c\dot{x}+kx=0$. L'équation caractéristique est $mz^2+cz+k=0$.

Oscillations non amorties : $c=0$. L'équation (h) s'écrit $\ddot{x}+(k/m)x=0$ et les solutions sont les fonctions $x(t)=\lambda_1\cos\omega t+\lambda_2\sin\omega t$, où $\omega=\sqrt{k/m}$. On peut aussi écrire les solutions sous la forme

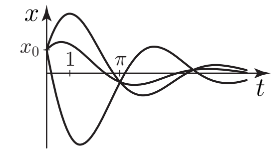
$$x(t)=a\cos(\omega t-\varphi),\quad \text{en posant } a=\sqrt{\lambda_1^2+\lambda_2^2} \text{ et } \tan\varphi=\lambda_2/\lambda_1.$$



Le mouvement est périodique, de période $T=2\pi/\omega$: pour une masse donnée, la fréquence $1/T$ augmente avec la raideur k du ressort. La figure montre des solutions pour un même déplacement initial $x(0)=x_0$ à partir de l'équilibre avec différentes vitesses initiales $\dot{x}(0)=v_0$, pente de la tangente en $t=0$.

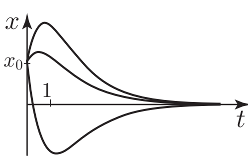
Oscillations amorties : $0<c^2<4mk$. Les racines de l'équation caractéristique sont les nombres complexes $-r\pm\boldsymbol{i}\omega$, où $r=c/(2m)$ et $\omega=(1/2m)\sqrt{4mk-c^2}$. Les solutions sont données par

$$x(t)=ae^{-rt}\cos(\omega t-\varphi),\ \text{où } a>0 \text{ et } \varphi \text{ sont des constantes.}$$

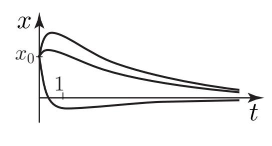


Le mouvement consiste en des oscillations d'amplitudes de plus en plus petites, les positions extrêmes étant limitées par les fonctions $\pm ae^{-rt}$ qui tendent vers 0 quand t tend vers $+\infty$: la masse revient en oscillant vers sa position d'équilibre.

Amortissement fort : $c^2 > 4mk$. L'équation caractéristique a deux racines réelles r_1, r_2 négatives (le produit $r_1 r_2 = k/m$ étant positif, les racines ont le même signe, celui de leur somme $r_1+r_2=-c/m<0$: voir page 52). Les solutions sont données par $x(t) = \lambda_1 e^{r_1 t} + \lambda_2 e^{r_2 t}$. La masse n'oscille pas, elle revient vers sa position d'équilibre $x = 0$ quand t tend vers l'infini.



Amortissement critique : $c^2 = 4mk$. L'équation caractéristique a une racine double $-c/2m$ et les solutions sont $x(t) = e^{-ct/2m}(\lambda_1 t + \lambda_2)$. La masse n'oscille pas. Puisqu'une exponentielle n'est jamais nulle, $x(t)$ s'annule exactement une fois : dans l'intervalle de temps $t \geqslant 0$, la masse ne peut passer qu'au plus une fois par sa position d'équilibre.



Dans certains circuits électroniques, les valeurs des composants sont calculées pour que le dispositif présente ce type de fonctionnement.

Recherche d'une solution de l'équation complète

Considérons un oscillateur mécanique sur lequel on fait agir une force variable $f(t)$, ou bien un circuit résistance-capacité-inductance alimenté par un générateur de tension variable $E(t)$. La *réponse* $x(t)$ à *l'excitation* est solution d'une équation différentielle $\ddot{x} + p\dot{x} + qx = b(t)$ de second membre $b(t) = f(t)$ ou $b(t) = E(t)$ selon le cas.
Pour trouver une solution particulière, on peut utiliser la formule générale donnée page 451. Mais plaçons-nous dans le cas fréquent d'un second membre de la forme

$$A(t)\cos \alpha t \text{ ou bien } A(t)\sin \alpha t, \text{ avec } A \text{ une fonction polynôme.}$$

On cherche alors directement une solution particulière

$$s(t) = U(t)\cos \alpha t + V(t)\sin \alpha t, \text{ où } U \text{ et } V \text{ sont des polynômes}$$

- de même degré que A si $\boldsymbol{i}\,\alpha$ n'est pas racine de l'équation caractéristique $z^2 + pz + q = 0$,
- de degré $1 + \deg A$ si $\boldsymbol{i}\,\alpha$ est racine de l'équation caractéristique.

Cas non amorti. Cherchons une solution de l'équation différentielle $\ddot{x} + \omega^2 x = E\cos \alpha t$ (E constant) sous la forme $s(t) = U(t)\cos \alpha t + V(t)\sin \alpha t$, où U et V sont des polynômes. L'équation caractéristique $z^2 + \omega^2 = 0$ a pour racines $\pm \boldsymbol{i}\,\omega$.

Supposons $\alpha \neq \pm\omega$. Les polynômes U et V sont alors des constantes. En posant $s_1(t) = U\cos \alpha t$ et $s_2(t) = V\sin \alpha t$, il vient

$$\ddot{s}_1(t) + \omega^2 s_1(t) = -\alpha^2 U\cos \alpha t + \omega^2 U\cos \alpha t \text{ , } \ddot{s}_2(t) + \omega^2 s_2(t) = -\alpha^2 V\sin \alpha t + \omega^2 V\sin \alpha t$$

La fonction $s(t)$ est solution si et seulement si l'on a

$$(\omega^2 - \alpha^2)U\cos \alpha t + (\omega^2 - \alpha^2)V\sin \alpha t = E\cos \alpha t \quad \text{quel que soit } t.$$

En identifiant, on obtient $V=0$, $U=\frac{E}{\omega^2-\alpha^2}$ et $s(t)=\frac{E\cos\alpha t}{\omega^2-\alpha^2}$.

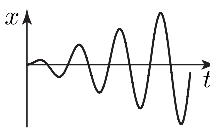
La solution générale $a\cos(\omega t-\varphi)+s(t)$ de l'équation est la superposition de deux régimes périodiques de fréquences $\omega/2\pi$ (la *fréquence propre*) et $\alpha/2\pi$ (la *fréquence d'excitation*).

Supposons $\alpha^2=\omega^2$. On cherche simplement un polynôme $U(t)=ut$ sans terme constant, car $\cos\alpha t$ est solution de l'équation homogène. En posant $s(t)=ut\sin\alpha t$, il vient $\dot{s}(t)=u\sin\alpha t+\alpha ut\cos\alpha t$ et

$$\ddot{s}(t)+\omega^2 s(t)=2u\alpha\cos\alpha t-u\alpha^2 t\sin\alpha t+\omega^2 ut\sin\alpha t=2u\alpha\cos\alpha t, \quad \text{car } \omega^2=\alpha^2.$$

La fonction $s(t)$ est donc solution si et seulement si $2u\alpha=E$, d'où la solution $s(t)=\frac{t\sin\alpha t}{2\alpha}E$.

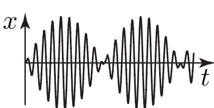
Cette solution particulière n'est pas bornée, elle présente des oscillations d'amplitudes croissantes, limitées par les droites de pente $\pm(1/2\alpha)$ issues de l'origine. On dit que l'excitation $E\cos\omega t$ provoque la *résonance* de l'oscillateur.



Phénomène de battement. Supposons $\alpha\neq\omega$, mais α et ω peu différents. La solution de conditions initiales $x(0)=\dot{x}(0)=0$ est

$$x(t)=\frac{E}{\alpha^2-\omega^2}\big[\cos\omega t-\cos\alpha t\big]=\frac{2E}{\alpha^2-\omega^2}\left[\sin\frac{(\alpha+\omega)t}{2}\sin\frac{(\alpha-\omega)t}{2}\right]$$

Puisque $|\alpha-\omega|$ est petit, la période de $\sin\frac{(\alpha-\omega)t}{2}$ est grande : les oscillations de $x(t)$ sont lentement modulées en amplitude.



Cas des oscillations amorties. Cherchons une solution particulière $s(t)$ de l'équation du pendule $m\ddot{x}+c\dot{x}+kx=E\cos\alpha t$, où les nombres positifs k, c et m vérifient $0<c^2<4mk$. En posant $s(t)=U\cos\alpha t+V\sin\alpha t$ et en identifiant dans l'équation, on obtient le système linéaire $\begin{cases}(k-m\alpha^2)U+c\alpha V &=E\\ -C\alpha U+(k-m\alpha^2)V &=0\end{cases}$ et la solution

$$U=\frac{E(k-m\alpha^2)}{(k-m\alpha^2)^2+c^2\alpha^2} \quad , \quad V=\frac{Ec\alpha}{(k-m\alpha^2)^2+c^2\alpha^2}$$

On a $s(t)=F\cos(\alpha t-\varphi)$, où $F=\sqrt{U^2+V^2}=\frac{E}{\sqrt{(k-m\alpha^2)^2+c^2\alpha^2}}$ est l'amplitude maximale de la fonction s.

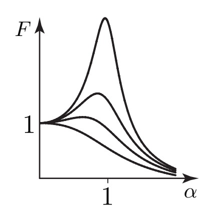
Quand t tend vers $+\infty$, chaque solution de l'équation homogène tend vers 0 comme une fonction e^{-rt}. Il s'ensuit que la réponse de l'oscillateur est rapidement très voisine de $s(t)$. Cette réponse est périodique, de même période que l'excitation, mais déphasée de φ, où $\tan\varphi=V/U=c\alpha/(k-m\alpha^2)$. Faisons varier la fréquence d'excitation et cherchons le comportement de l'amplitude F.

Supposons $\alpha>0$.

Comme fonction de α, F varie dans le sens contraire de $(k-m\alpha^2)^2+c^2\alpha^2$. La dérivée $\frac{dF}{d\alpha}$ a donc le signe de $-\frac{d}{d\alpha}\big[(k-m\alpha^2)^2+c^2\alpha^2\big]=-2\alpha\big[(c^2-2mk)+2m^2\alpha^2\big]$.

► Si $c^2 > 2mk$, la fonction $F(\alpha)$ est décroissante.

► Supposons $0 < c^2 \leqslant 2mk$. Alors $F(\alpha)$ est maximum pour une certaine valeur α_0 telle que $2m^2\alpha_0^2 = 2mk - c^2$. Le quotient F/E est un gain d'amplitude entre l'excitation et la réponse : s'il est supérieur à 1, l'oscillateur fonctionne en amplificateur.

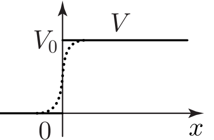


Sur la figure, nous avons porté le gain en ordonnée et α en abscisse. Chaque courbe montre les variations du gain pour une valeur donnée de c ; le gain maximum est d'autant plus grand que c est plus petit. Quand le coefficient d'amortissement c tend vers 0, le gain maximum tend vers l'infini et α_0 tend vers $\sqrt{k/m}$ qui est la fréquence d'excitation provoquant la résonance dans l'équation $m\ddot{x} + kx = 0$ (page 456).

Application : franchissement d'une marche de potentiel

Considérons le mouvement rectiligne d'une particule d'énergie constante E dans un champ de potentiel $V(x)$, où x représente l'abscisse sur la trajectoire. En Mécanique quantique, sa fonction d'onde $\varphi(x)e^{-i\omega t}$ doit (dans un état stationnaire) vérifier l'équation différentielle $\dfrac{d^2\varphi}{dx^2} + \dfrac{2m}{\hbar^2}(E-V)\varphi = 0$, où $\hbar$ est la constante de Planck et m la masse de la particule.

Prenons le potentiel V défini par $V(x) = 0$ si $x < 0$, $V(x) = V_0$ si $x > 0$, et supposons $V_0 > 0$: on dit que c'est une « marche de potentiel » ; cette fonction discontinue en 0 modélise un potentiel physique (en pointillé sur la figure) qui varie continûment de 0 à V_0 sur un très petit intervalle de longueur négligeable devant la longueur d'onde de la particule. L'équation différentielle s'écrit



$$\frac{d^2\varphi}{dx^2} + \frac{2mE}{\hbar^2}\varphi = 0 \quad \text{dans le domaine } x < 0$$
$$\frac{d^2\varphi}{dx^2} + \frac{2m(E-V_0)}{\hbar^2}\varphi = 0 \quad \text{dans le domaine } x > 0$$

Cas $E > V_0$: réflexion partielle. En Mécanique classique, une particule d'énergie E franchit toujours une marche de potentiel $V_0 < E$. Nous allons voir que, dans le cadre de la Mécanique quantique, l'onde associée a une probabilité non nulle d'être réfléchie.

Posons $k_1 = \sqrt{2mE}/\hbar$ et $k_2 = \sqrt{2m(E-V_0)}/\hbar$. Sous forme complexe, la solution s'écrit

$$\varphi(x) = \begin{cases} \varphi_1(x) = a_1 e^{i k_1 x} + a_1' e^{-i k_1 x} & \text{si } x < 0 \\ \varphi_2(x) = a_2 e^{i k_2 x} + a_2' e^{-i k_2 x} & \text{si } x > 0 \end{cases}$$

On doit calculer les constantes pour raccorder φ_1 et φ_2 en $x = 0$ de sorte que le résultat soit continu et dérivable en ce point (puisque $V(x)$ n'est pas continu en $x = 0$, l'équation différentielle montre que les solutions ne peuvent pas avoir de dérivée seconde en ce point).

Derrière la barrière de potentiel, il n'y a pas de particule progressant dans le sens des abscisses décroissantes, donc $a'_2 = 0$: il n'y a qu'une seule onde transmise. Les conditions de raccordement $\varphi_1(0) = \varphi_2(0)$ et $\varphi'_1(0) = \varphi'_2(0)$ s'écrivent $a_1 + a'_1 = a_2$ et $k_1(a_1 - a'_1) = k_2 a_2$. Puisque la solution n'est déterminée qu'à un facteur multiplicatif près, introduisons les rapports $r = a'_1/a_1$ et $\lambda = a_2/a_1$. Il vient

$$\begin{cases} \lambda - r &= 1 \\ k_2\lambda + k_1 r &= k_1 \end{cases} \quad \text{d'où } r = \frac{k_1 - k_2}{k_1 + k_2} \quad \text{et} \quad \lambda = 1 + r = \frac{2k_1}{k_1 + k_2}.$$

Il existe donc des solutions $a_2 e^{\boldsymbol{i}\, k_2 x}$ définies pour $x > 0$, autrement dit l'onde incidente peut franchir la marche de potentiel.
Dans la fonction φ_1, la partie $a_1 e^{\boldsymbol{i}\, k_1 x}$ correspond à une onde progressant dans le sens x croissant avec une probabilité de présence uniforme proportionnelle à $|a_1 e^{\boldsymbol{i}\, k_1 x}|^2 = a_1^2$; la partie $a'_1 e^{-\boldsymbol{i}\, k_1 x}$ correspond à une onde dans sens opposé, de probabilité de présence ${a'_1}^2$: le *taux de réflexion* est $R = \left[\dfrac{a'_1}{a_1}\right]^2$. Le *taux de transmission* est évidemment $T = 1 - R$, car il est certain que la particule est transmise ou réfléchie. On a $R = r^2 = \dfrac{(k_1 - k_2)^2}{(k_1 + k_2)^2} = 1 - \dfrac{4k_1 k_2}{(k_1 + k_2)^2}$ et donc $T = \dfrac{4k_1 k_2}{(k_1 + k_2)^2}$.

Puisque les rapports $r = a'_1/a_1$ et $\lambda = a_2/a_1$ sont réels, la réflexion et la transmission se font sans déphasage. Si E est très grand par rapport à V_0, alors k_1 et k_2 sont proches, T est voisin de 1 et la particule est presque sûrement transmise.

Cas $E < V_0$: réflexion totale. En Mécanique classique, la particule ne peut pas franchir la marche de potentiel : elle est réfléchie. Voyons ce qu'il en est dans le cadre quantique.
On pose cette fois $\rho_2 = \sqrt{2m(V_0 - E)}/\hbar$ et l'on a $\varphi_2(x) = b_2 e^{\rho_2 x} + b'_2 e^{-\rho_2 x}$, pour $x > 0$. La solution doit rester bornée quand x tend vers $+\infty$, donc $b_2 = 0$. Les conditions de raccordement s'écrivent

$$\frac{a'_1}{a_1} = \frac{k_1 - \boldsymbol{i}\,\rho_2}{k_1 + \boldsymbol{i}\,\rho_2} \quad \text{et} \quad \frac{b'_2}{a_1} = \frac{2k_1}{k_1 + \boldsymbol{i}\,\rho_2}$$

et le taux de réflexion est $R = \left[\dfrac{a'_1}{a_1}\right]^2 = \left|\dfrac{k_1 - \boldsymbol{i}\,\rho_2}{k_1 + \boldsymbol{i}\,\rho_2}\right| = 1$. Comme en Mécanique classique, l'onde est donc réfléchie et puisque le terme a'_1/a_1 n'est pas réel, l'onde réfléchie est déphasée par rapport à l'onde incidente. Cependant, la solution réelle $b'_2 e^{-\rho_2 x}$ définie pour $x > 0$, montre que la particule a une probabilité non nulle d'exister au delà de la marche de potentiel, contrairement au cas classique. L'onde transmise est qualifiée d'évanescente, car $b'_2 e^{-\rho_2 x}$ tend vers 0 quand x tend vers $+\infty$.

2.3 Conditions aux bords

Pour déterminer une solution d'une équation différentielle du second ordre, nous avons jusqu'à présent utilisé des conditions initiales $x(t_0) = x_0$, $x'(t_0) = v_0$ portant sur la valeur de la solution et de sa dérivée à un instant t_0 donné (théorème page 450).

Mais dans de nombreux problèmes, on cherche plutôt une solution prenant des valeurs données x_0 et x_1 à des instants t_0 et t_1 donnés. Ces conditions s'écrivent $x(t_0) = x_0$, $x(t_1) = x_1$, avec $t_0 \neq t_1$; comme on impose les valeurs de la solution aux extrémités de l'intervalle $[t_0, t_1]$, on dit que ce sont des *conditions aux bords.*
Nous allons voir que le problème n'a pas toujours de solution.

Exemple. L'équation différentielle $x'' + \omega^2 x = 0$ (avec $\omega \neq 0$) a pour solutions $x(t) = a\cos\omega t + b\sin\omega t$. Les solutions telles que $x(0) = x_0$ sont les fonctions

$$x(t) = x_0 \cos\omega t + b\sin\omega t$$

Donnons-nous un instant $t_1 \neq 0$ et une valeur x_1. La condition $x(t_1) = x_1$ s'écrit $x_0 \cos\omega t_1 + b\sin\omega t_1 = x_1$, ce qui permet de calculer b si ωt_1 n'est pas multiple entier de π.
Supposons $t_1 = \pi/\omega$, donc $x(t_1) = x_0 \cos\omega t_1 = -x_0$.

- Si $x_1 \neq -x_0$, il n'y a aucune solution vérifiant les conditions $x(0) = x_0$, $x(t_1) = x_1$.
- Si $x_1 = -x_0$, il y a une infinité de solutions telle que $x(0) = x_0$, $x(t_1) = x_1$: ce sont toutes les fonctions $x(t) = x_0 \cos\omega t + b\sin\omega t$, où b est quelconque.

Pour une équation différentielle linéaire à coefficients constants, on sait trouver toutes les solutions : il est donc possible de chercher s'il y a une solution vérifiant des conditions données aux bords, et de la calculer.

Cas d'une équation à coefficients variables

Abordons le problème des conditions aux bords pour une équation différentielle à coefficients variables :

$$(c) \qquad x'' + p(t)x' + q(t)x = b(t)\ , \quad x(t_0) = x_0\ ,\ x(t_1) = x_1$$

Tir au but. Considérons les solutions de l'équation différentielle qui prennent la valeur x_0 en t_0. On sait que chacune d'elles est déterminée par sa pente initiale $x'(t_0)$. En notant $s_k(t)$ la solution telle que $s'_k(t_0) = k$, le problème est de trouver la valeur de k pour que $s_k(t_1) = x_1$.
Pour résoudre numériquement le problème (c), on emploie d'habitude la technique du « tir au but ».

- Par balayage, on cherche d'abord une pente initiale k pour laquelle $s_k(t_1)$ n'est pas trop différent de x_1.
- Si par exemple $s_k(t_1) < x_1$, on fait varier k par petits pas dans le sens qui fait augmenter la valeur $s_k(t_1)$. Quand $s_k(t_1)$ a dépassé la valeur x_1, on fait évoluer k dans l'autre sens en employant un pas plus petit, et l'on itère ce processus d'approximations successives.

Bien entendu, on doit disposer d'une méthode numérique pour calculer la valeur en t_1 de la solution s_k de conditions initiales $s_k(t_0) = x_0$, $s'_k(t_0) = k$ (voir page 518).

s
x_1
x_0
tir trop haut
tir trop bas
t_0
t_1 t

Un résultat général

- *Soit* $u(t)$ *la solution de l'équation* (e) $\; x'' + p(t)x' + q(t)x = b(t)$ *telle que* $u(t_0) = x_0$ *et* $u'(t_0) = 0$.
- *Soit* $v(t)$ *la solution de l'équation homogène* (h) $\; x'' + p(t)x' + q(t)x = 0$ *telle que* $v(t_0) = 0$ *et* $v'(t_0) = 1$.

Si $v(t_1) \neq 0$*, la fonction* $x(t) = u(t) + \dfrac{x_1 - u(t_1)}{v(t_1)} v(t)$ *est solution de* (e) *et satisfait les conditions aux bords* $x(t_0) = x_0$*,* $x(t_1) = x_1$*.*

Démonstration. La fonction $u(t)$ est une solution de (e) et $\lambda v(t)$ est solution de (h) pour toute constante λ, donc $x(t)$ est solution de (e). De plus, on a $x(t_0) = u(t_0) + 0 = x_0$ et $x(t_1) = u(t_1) + \dfrac{x_1 - u(t_1)}{v(t_1)} v(t_1) = u(t_1) + x_1 - u(t_1) = x_1$, donc $x(t)$ satisfait bien les conditions aux bords. ■

3. L'équation de Newton

En Mécanique, on considère souvent des mouvements dont l'accélération n'est fonction que de la position : c'est le cas, par exemple, pour une masse oscillant sans amortissement au bout d'un ressort (page 454). Étudions l'équation différentielle générale de ce type de mouvement.

> **Définition**
> Une *équation de Newton* est une équation différentielle de la forme $x'' = f(x)$, où f est une fonction continue sur un intervalle.

Résolution

On commence par chercher s'il y a des *équilibres*, c'est-à-dire des solutions $x(t)$ constantes.

Si a est un nombre tel que $f(a) = 0$, alors la fonction constante $x(t) = a$ est solution, car $x'(t) = x''(t) = 0 = f\big(x(t)\big)$ quel que soit t. Réciproquement, pour qu'une fonction constante $x(t) = a$ soit solution, il faut que l'on ait $0 = x''(t) = f\big(x(t)\big) = f(a)$.

Les valeurs d'équilibre sont les nombres a *tels que* $f(a) = 0$ *:*
la fonction constante de valeur a *est solution.*

Si a est une valeur d'équilibre, la solution de conditions initiales $x(t_0) = a, x'(t_0) = 0$ est la fonction constante de valeur a. Une fois trouvés les équilibres, on résout l'équation dans un domaine ne contenant pas d'équilibre.

Première étape. En multipliant par x', l'équation s'écrit $x'x'' = f(x)x'$.
Posons $v = x'$ et soit F une primitive de f.
On a $\dfrac{1}{2}\dfrac{d}{dt}(v^2) = vv' = x'x''$ et $\dfrac{d}{dt}\big[F\big(x(t)\big)\big] = f\big(x(t)\big)x'(t) = x''(t)x'(t)$, donc

$$\frac{1}{2}\frac{d}{dt}(v^2) = \frac{d}{dt}\big[F\big(x(t)\big)\big]$$

Donnons-nous des conditions initiales $x(t_0) = x_0$ et $x'(t_0) = v_0$. En intégrant de t_0 à t l'égalité précédente, il vient

(1) $$\frac{1}{2}v^2 - \frac{1}{2}v_0^2 = F(x) - F(x_0)$$

L'égalité (1) s'appelle une *intégrale première* de l'équation, car elle ne fait plus intervenir que la dérivée première $v = x'$.

Seconde étape. On résout l'intégrale première. L'égalité (1) s'écrit en effet $x'^2 = 2F(x) - 2F(x_0) + v_0^2$, ou encore

(2) $$x' = \pm\sqrt{2F(x) - 2F(x_0) + v_0^2}$$

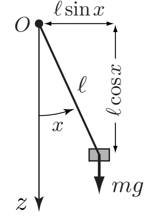
C'est une équation différentielle autonome du premier ordre (page 447) dont il faut calculer la solution de condition initiale $x(t_0) = x_0$.

L'exemple du pendule

Faisons osciller une masse m suspendue en un point O par une tige non pesante de longueur ℓ. On suppose que le mouvement se fait dans un plan vertical passant par O.

Considérons un axe Oz dirigé vers le bas par un vecteur unitaire $\vec{u}$ et appelons x l'angle que fait la tige avec cet axe.
La masse est soumise à son poids $mg\vec{u}$ dont le moment de rappel par rapport à O est $\mathscr{M} = -mg\ell \sin x$. Le moment d'inertie de la masse par rapport à O étant $I = m\ell^2$, le mouvement est régi par l'équation $\mathscr{M} = I\ddot{x}$.
Après simplification par $m\ell$, cette égalité s'écrit $-g \sin x = \ell\ddot{x}$, c'est-à-dire

(*e*) $$\ddot{x} = -(g/\ell)\sin x$$



C'est une équation de Newton, avec $f(x) = -(g/\ell)\sin x$. Remarquons qu'elle ne dépend pas de la masse du pendule.
Les valeurs d'équilibre sont les $a_k = k\pi$, où k est un nombre entier quelconque. Il y a deux positions d'équilibre : pour les valeurs $0, 2\pi, \ldots$, le pendule est en position verticale basse ; pour les valeurs $\pi, 3\pi, \ldots$, il est en position verticale haute. Si l'on place le pendule dans l'une de ces positions et qu'on l'abandonne sans vitesse initiale, il reste en équilibre.

Supposons qu'on écarte le pendule d'un angle x_0 et qu'on le lâche avec une vitesse angulaire initiale v_0.

Calcul de l'intégrale première. En prenant $\cos x$ comme primitive de $-\sin x$, l'intégrale première est $v^2 - v_0^2 = 2(g/\ell)\cos x - 2(g/\ell)\cos x_0$, ou encore

$$v^2 - 2(g/\ell)\cos x = v_0^2 - 2(g/\ell)\cos x_0$$

À l'instant initial, la vitesse de la masse est ℓv_0, donc son énergie cinétique est $E_c = \frac{1}{2}m\ell^2 v_0^2$. La hauteur de la masse par rapport au point O est $-\ell\cos x_0$, donc la différence d'énergie potentielle par rapport à ce point est $E_p = -mg\ell\cos x_0$. L'énergie totale est $U = E_c + E_p = m\ell^2\left[\frac{1}{2}v_0^2 - (g/\ell)\cos x_0\right] = m\ell^2 K_0$. Puisque le système n'est soumis qu'à la pesanteur, la loi de conservation de l'énergie (page 425) affirme que U

reste constante au cours du mouvement : c'est précisément ce qu'exprime l'intégrale première, sous la forme $v^2-2(g/\ell)\cos x = 2K_0$.

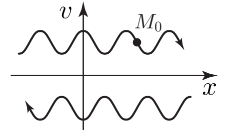
D'après l'intégrale première, la solution $x(t)$ de conditions initiales $x(0)=x_0$, $\dot{x}(0)=v_0$ est déterminée par $\dot{x}^2 = 2(g/\ell)\cos x + 2K_0$, où $K_0 = \dfrac{v_0^2}{2} - (g/\ell)\cos x_0$.
En posant $L_0 = (\ell/g)K_0$, cette équation différentielle s'écrit

(1) $$x(0) = x_0 \quad , \quad \dot{x}^2 = 2(g/\ell)(L_0 + \cos x)$$

Son domaine est formé des (t,x) tels que $L_0+\cos x \geqslant 0$. Si $L_0 \geqslant 1$, alors on a $L_0+\cos x \geqslant 0$ pour tout x et le domaine est l'ensemble de tous les couples (t,x). Si $-1<L_0<1$, alors x doit rester dans les intervalles où $\cos x \geqslant -L_0$. On a toujours $L_0=(\ell/g)K_0 \geqslant -\cos x_0 \geqslant -1$.

Le plan des phases. Repérons le mouvement de la masse par sa position x et sa vitesse angulaire $\dot{x}=v$. Portons x en abscisse et v en ordonnée : on obtient le *plan des phases*. À tout instant t, l'état du pendule correspond à un point $M(t) = \big(x(t), v(t)\big)$ dans le plan des phases.
Le mouvement est donc représenté par une courbe paramétrée dans ce plan. Pour un point initial $M_0 = M(0)$ donné, l'ensemble des points $M(t)$, $t \in \mathbb{R}$, s'appelle la *trajectoire* de M_0.
Dans le demi-plan supérieur, on a $\dot{x} = v > 0$, donc la trajectoire est parcourue dans le sens x croissant ; dans le demi-plan inférieur, la trajectoire est parcourue dans le sens x décroissant. D'après l'intégrale première, le point $M(t)$ est toujours sur la courbe C d'équation

$$v^2 = 2(g/\ell)(L_0 + \cos x)$$



Étude des trajectoires dans le plan des phases. On part du graphe de $\cos x$; en translatant l'axe des x de la quantité $-L_0$ le long de l'axe des ordonnées, on obtient le graphe de $L_0+\cos x$. Pour dessiner la courbe C, étudions la fonction

$$h(x) = \sqrt{2g/\ell}\,\sqrt{L_0+\cos x}\,.$$

La courbe C n'est définie que pour les valeurs de x où l'on a $L_0+\cos x \geqslant 0$. Procédons comme page 275.

Cas $L_0 > 1$. On a $L_0+\cos x > 0$ pour tout x, donc h est définie sur $\mathbb{R}$, $h(x)$ est strictement positif et varie comme $\cos x$. Puisque $v(x) = \pm h(x)$, C est la réunion du graphe de h (courbe C_+) et du graphe de $-h$ (courbe C_-), deux courbes sans point commun (figure 1).

Supposons $v_0 > 0$. Le point initial $M_0 = (x_0, v_0)$ est donc sur la courbe C_+ située dans le demi-plan $v > 0$ et $M(t)$ parcourt C_+ dans le sens x croissant. De plus, $\dot{x}(t)$ reste supérieur ou égal au nombre positif $\sqrt{(2g/\ell)(L_0-1)}$, donc $x(t)$ tend vers $+\infty$ quand t tend vers $+\infty$: la trajectoire est la courbe C_+ tout entière. La vitesse est maximum pour $x = 2k\pi$ (passage du pendule en position basse) et minimum pour $x = (2k+1)\pi$. Pour le pendule, il s'agit d'un mouvement tournoyant autour du point d'attache O.

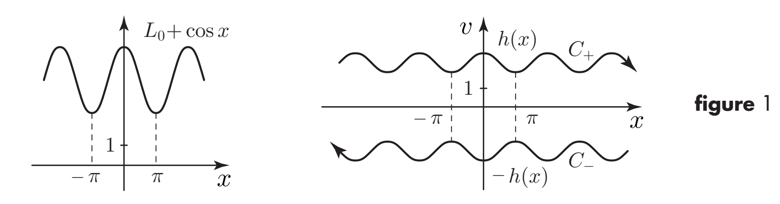


figure 1

Si $v_0 < 0$, on obtient le mouvement en sens inverse : la trajectoire est la courbe C_- située dans le demi-plan $v < 0$, parcourue dans le sens x décroissant.

Ces mouvements ne peuvent exister qu'à un niveau d'énergie suffisant et avec une vitesse initiale v_0 non nulle.

Cas $-1 < L_0 < 1$. Il y a un nombre $\alpha \in]0, \pi[$ tel que $\cos\alpha = -L_0$. Puisqu'on a $\cos x \geqslant \cos\alpha$ si $-\alpha \leqslant x \leqslant \alpha$, la fonction $h(x)$ est définie sur $I = [-\alpha, \alpha]$ et aussi sur les intervalles qui s'en déduisent par translation de 2π. Sur I, $h(x)$ varie comme $\cos x$ et s'annule en α et en $-\alpha$ avec une tangente verticale (page 276). Pour avoir la partie C_0 de C située entre les abscisses $-\alpha$ et α, on complète le graphe de h en faisant la symétrie par rapport à l'axe des abscisses, ce qui donne une courbe fermée (figure 2). La fonction cosinus étant paire, C_0 est aussi symétrique par rapport à l'axe des ordonnées. La courbe C est formée de C_0 et de toutes ses translatées de $2k\pi$.

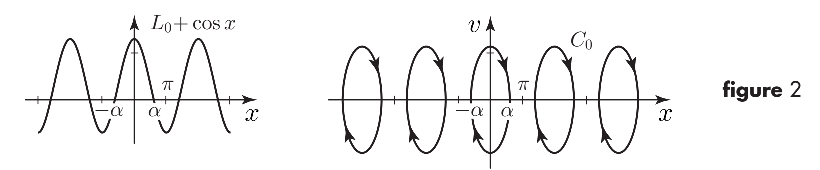


figure 2

Pour décrire le mouvement, on peut supposer x_0 entre $-\pi$ et π : le point initial (x_0, v_0) étant sur C_0, $M(t)$ parcourt le cycle C_0 indéfiniment. Le sens est celui des x croissants quand $M(t)$ se trouve dans le demi-plan $v > 0$, et celui des x décroissants quand le point est dans le demi-plan $v < 0$. L'angle x et la vitesse v sont périodiques : le mouvement consiste en des oscillations d'amplitude α autour de la position d'équilibre basse.

Cas $L_0 = 1$. Pour $-\pi \leqslant x \leqslant \pi$, on a $h(x) = \sqrt{2g/\ell}\sqrt{1 + \cos x} = 2\sqrt{g/\ell}\,\cos(x/2)$, car $1 + \cos x = 2\big(\cos(x/2)\big)^2$ et $\cos(x/2) \geqslant 0$.

Sur $[-\pi, \pi]$, on a $h'(x) = -\sqrt{g/\ell}\,\sin(x/2)$: en π, h a une demi-tangente de pente $-\sqrt{g/\ell}$, et en $-\pi$, h a une demi-tangente de pente opposée $\sqrt{g/\ell}$. On en déduit le graphe de h et, par symétrie, la courbe C (figure 3).

La différence par rapport au cas précédent, c'est que les points d'équilibre $A = (-\pi, 0)$ et $B = (\pi, 0)$ sont sur C. Notons T^+ la partie de C située dans le demi-plan $v > 0$

et T^- la partie située dans le demi-plan $v < 0$.

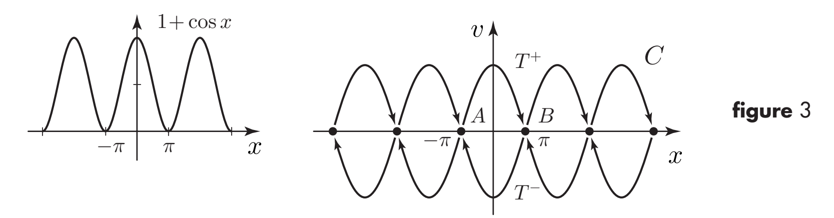


figure 3

► Si le point initial est A, c'est-à-dire $x_0 = -\pi$ et $v_0 = 0$, la trajectoire est réduite au point A : le pendule reste dans sa position d'équilibre haute ; de même si le point initial est en B, la trajectoire est réduite au point B.

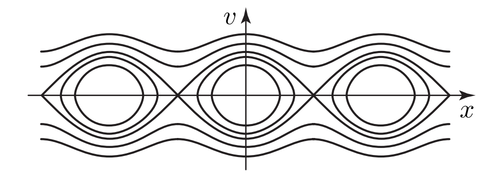
► Supposons que le point initial M_0 est sur T^+, autrement dit $-\pi < x_0 < \pi$ et $v_0 > 0$. La trajectoire est alors tout l'arc $\overset{\frown}{AB}$ de T^+, extrémités non comprises, parcouru de A vers B. Le point $M(t)$ met un temps infini pour parcourir l'arc $\overset{\frown}{M_0B}$.

► Si l'on a $-\pi < x_0 < \pi$ et $v_0 < 0$, le point initial est sur T^- et la trajectoire est l'arc $\overset{\frown}{BA}$ de T^-, extrémités non comprises, parcouru de B vers A.

Il y a donc quatre trajectoires entre $-\pi$ et π : les deux points A et B et les arcs T^+ et T^-.

Cas $L_0 = -1$. Puisque $-1+\cos x$ n'est positif ou nul que pour $x = 2k\pi$, où k est un entier, l'ensemble C n'est constitué que des points d'équilibre bas $(2k\pi, 0)$, chacun d'eux étant une trajectoire.

La figure ci-dessous montre l'ensemble des trajectoires.



Pour déterminer la fonction $x(t)$, il faudrait résoudre l'équation différentielle (1) page 462. En écrivant $\dfrac{dx}{\sqrt{L_0+\cos x}} = \pm\sqrt{2g/\ell}\;dt$, il vient

$$\text{(2)} \qquad \int_{x_0}^{x(t)} \frac{du}{\sqrt{L_0+\cos u}} = \pm\sqrt{\frac{2g}{\ell}}\;t$$

Mais l'intégrale ne se calcule pas au moyen des fonctions usuelles.

Stabilité des équilibres

► L'équilibre bas $x_0 = v_0 = 0$ est *stable* : si M_0 est proche de l'origine, la trajectoire est un petit cycle autour de l'origine, correspondant aux petites oscillations du pendule autour de son équilibre bas.

► L'équilibre haut $x_0=\pm\pi, v_0=0$ est instable : en effet, si M_0 est proche de $B=(\pi,0)$ mais différent de B, le point $M(t)$ parcourt une trajectoire qui l'éloigne loin de B ; c'est ce qui se passe si l'on abandonne le pendule sans vitesse initiale à partir d'une position angulaire x_0 voisine de π mais différente de π (figure 2 avec α proche de π), ou bien si, à partir de la position $x_0=\pm\pi$, on lui donne une vitesse initiale v_0 non nulle (figure 1).

Étude des petites oscillations sans vitesse initiale. Supposons qu'on lâche le pendule sans vitesse initiale ($v_0=0$) après l'avoir écarté d'un angle x_0 tel que $0<x_0<\pi$.

En posant $\sqrt{g/\ell}=\omega$, on a alors $K_0=-\omega^2\cos x_0$ et $L_0=-\cos x_0$. Pendant la première demi-oscillation, on a $-x_0\leqslant x(t)\leqslant x_0$, $\dot{x}(t)<0$ et l'équation (2) devient

(3) $$\frac{dx}{\sqrt{\cos x-\cos x_0}}=-\sqrt{2}\,\omega\,dt$$

Supposons x_0 petit.

Remplaçons $\cos x$ par son développement limité $1-x^2/2$ en 0 et de même pour $\cos x_0$. On obtient $\dfrac{dx}{\sqrt{x_0^2-x^2}}=-\omega\,dt$, et en posant $x=x_0y$, il vient $\dfrac{dy}{\sqrt{1-y^2}}=-\omega dt$.

En intégrant, on a $\operatorname{Arc\,sin} y=-\omega t+c$, d'où $y=\sin(-\omega t+c)$; la constante c est déterminée par $y=1$ quand $t=0$, donc $1=\sin c$, $c=\pi/2$ et $y=\sin(-\omega t+\pi/2)=\cos\omega t$.

Finalement, quand x_0 est très petit, les oscillations sont approximativement décrites par la fonction

$$x(t)=x_0\cos\omega t$$

La période est $T_0=2\pi/\omega=2\pi\sqrt{\ell/g}$, indépendante de la petite amplitude x_0.

En réalité, la période T dépend de x_0. Pour en faire une meilleure estimation, revenons à l'égalité (3). Le temps mis par le pendule pour aller de x_0 à 0 est un quart de période, donc d'après (3), on a

$$\sqrt{2}\,\omega\frac{T}{4}=\int_0^{x_0}\frac{dx}{\left(\cos x-\cos x_0\right)^{1/2}}$$

Cette intégrale généralisée est bien définie, comme nous l'avons montré page 327.

Puisque $\cos x-\cos x_0=2\sin^2(x_0/2)-2\sin^2(x/2)$, il vient

$$\frac{T}{4}\frac{2\pi}{T_0}\sqrt{2}=\frac{1}{\sqrt{2}}\int_0^{x_0}\frac{dx}{\left[\sin^2(x_0/2)-\sin^2(x/2)\right]^{1/2}}$$

ou encore

$$\pi\frac{T}{T_0}=\int_0^{x_0}\frac{dx}{\left[\sin^2(x_0/2)-\sin^2(x/2)\right]^{1/2}}$$

Puisque x reste inférieur à x_0, faisons le changement de variable $\sin(x/2)=\sin(x_0/2)\sin\varphi$. On a

$$\frac{1}{2}\cos\frac{x}{2}\,dx=\sin\frac{x_0}{2}\cos\varphi\,d\varphi$$

$$\pi\frac{T}{T_0} = \int_0^{\pi/2} \frac{2\sin(x_0/2)\cos\varphi\,d\varphi}{\cos(x/2)\left[\sin^2(x_0/2) - \sin^2(x_0/2)\sin^2\varphi\right]^{1/2}}$$
$$= 2\int_0^{\pi/2} \frac{\sin(x_0/2)\cos\varphi\,d\varphi}{\cos(x/2)\sin(x_0/2)\cos\varphi}$$
$$\pi\frac{T}{T_0} = 2\int_0^{\pi/2} \frac{d\varphi}{\cos(x/2)} = 2\int_0^{\pi/2} \frac{d\varphi}{\left[1 - \sin^2(x_0/2)\sin^2\varphi\right]^{1/2}}$$

Puisque x_0 est petit, remplaçons la fraction par son développement limité :

$$\frac{\pi}{2}\frac{T}{T_0} \sim \int_0^{\pi/2} \left[1 + \frac{1}{2}\sin^2(x_0/2)\sin^2\varphi\right] d\varphi$$
$$\sim \int_0^{\pi/2} \left[1 + \frac{x_0^2}{16}(1 - \cos 2\varphi)\right] d\varphi \ , \quad \text{car } \sin(x_0/2) \simeq x_0/2$$
$$\sim \left[1 + \frac{x_0^2}{16}\right]\frac{\pi}{2} \ , \qquad \text{car } \int_0^{\pi/2} \cos 2\varphi\,d\varphi = 0$$

On trouve finalement que la période est $T \sim T_0\left(1 + \dfrac{x_0^2}{16}\right)$ quand x_0 est petit.

4. Introduction au calcul des variations

De nombreux problèmes conduisent à chercher une fonction $y(x)$ qui rende maximum (ou minimum) une quantité de la forme $J(y) = \int_a^b F\big(x, y(x), y'(x)\big)\,dx$, où l'expression sous le signe intégrale dépend de x, de la fonction y et de sa dérivée y'. L'intégrale $J(y)$ s'appelle une *fonctionnelle de* y.
En général, la fonction inconnue y doit satisfaire des conditions supplémentaires, comme prendre des valeurs données en $x = a$ et en $x = b$ (conditions aux extrémités).

Exemple. La vitesse de la lumière dans un matériau d'indice optique n est c/n, où c est la vitesse de la lumière dans le vide. D'après le principe de Fermat, un rayon lumineux suit le chemin le plus rapide. Pour un déplacement suivant une courbe $y(x)$ dans la direction x croissant, un élément d'arc de longueur $ds = \sqrt{1 + y'^2}$ est parcouru à la vitesse $c/n(x)$, où $n(x)$ est l'indice du milieu à l'abscisse x. On a donc $\dfrac{ds}{dt} = \dfrac{c}{n(x)}$ et le temps mis par la lumière pour aller d'un point A d'abscisse a à un point B d'abscisse b est l'intégrale de $c\,dt$, c'est-à-dire la fonctionnelle

$$J(y) = \frac{1}{c}\int_a^b n(x)\,\sqrt{1 + y'^2}\,dx$$

Énoncé du problème. Pour clarifier la notation, introduisons une variable p pour désigner la dérivée y'. Considérons une fonction $F(x, y, p)$ de trois variables et la

fonctionnelle

$$J(y) = \int_a^b F\big(x, y(x), y'(x)\big)\, dx$$

Les fonctions admissibles. Donnons-nous des points $A = (a, y_A)$ et $B = (b, y_B)$: on dira qu'une fonction $y(x)$, dérivable jusqu'à un ordre suffisant, est *admissible* si l'on a $y(a) = y_A$ et $y(b) = y_B$, autrement dit si le graphe de la fonction $y(x)$ a pour extrémités les points A et B.
Dans le cas où l'on veut rendre maximum la fonctionnelle $J(y)$, le problème se formule ainsi :

trouver une fonction $f(x)$ admissible telle que $J(f) \geqslant J(y)$ pour toute fonction admissible y.

Variation d'une fonctionnelle. Puisque y est une fonction, un accroissement h à la fonction y est une fonction $h(x)$ (suffisamment de fois dérivable) telle que la fonction $(y+h)(x) = y(x) + h(x)$ est encore admissible, ce qui revient à supposer $h(a) = h(b) = 0$.

Si l'on ajoute deux accroissements ou si l'on multiplie un accroissement par un nombre réel, ces conditions aux extrémités sont encore vérifiées : l'ensemble des accroissements est donc un espace vectoriel de fonctions. Afin de mesurer la taille d'un accroissement h, définissons sa norme

$$\|h\| = \max_{a \leqslant x \leqslant b} \big[|h(x)| + |h'(x)|\big]$$

Pour cette norme, un accroissement h est petit si toutes ses valeurs $h(x)$, ainsi que les valeurs de sa dérivée $h'(x)$, sont petites.

Supposons qu'on peut écrire

$$J(y+h) - J(y) = L_y(h) + \varphi(h)$$

où $L_y(h)$ dépend linéairement de h et $\varphi(h) \underset{h \to 0}{\ll} \|h\|$ (donc $\varphi(h)$ est négligeable devant $\|h\|$ quand h tend vers 0, voir page 265). Pour un accroissement h assez petit, la fonctionnelle $L_y(h)$ ne diffère de la différence $J(y+h) - J(y)$ que d'une quantité négligeable devant $\|h\|$.
La fonctionnelle linéaire $L_y(h)$ s'appelle la *variation de J en y*.

La variation d'une fonctionnelle est l'analogue de la différentielle ordinaire d'une fonction. Comme pour la différentielle, nous allons voir que si $J(y)$ présente un extremum en $y = f$, alors sa variation en f est nulle.

Condition nécessaire d'extremum. *Soit f une fonction admissible. Si la fonctionnelle $J(y)$ présente un extremum local en $y = f$, alors $L_f(h) = 0$ pour tout accroissement h tel que $h(a) = h(b) = 0$.*

Démonstration. Supposons par exemple que $J(y)$ a un maximum local en $y = f$, donc $J(f+h) - J(f) \leqslant 0$ pour tout h assez petit. Pour montrer que $L_f = 0$, raisonnons par l'absurde en supposant qu'il y a un accroissement h_0 tel que $L_f(h_0) \neq 0$. Puisque $J(f+h_0) - J(f)$ ne diffère de $L_f(h_0)$ que d'une quantité négligeable devant $\|h_0\|$, on en déduit, si la norme de

h_0 est assez petite, que $J(f+h_0)-J(f)$ et $L_f(h_0)$ sont tous deux négatifs. Mais puisque $L_f(h)$ dépend linéairement de h, on a aussi $L_f(-h_0)=-L_f(h_0)>0$ et comme $L_f(-h_0)$ doit avoir aussi le signe de $J(f-h_0)-J(f)\leqslant 0$, cela est impossible. ■

4.1 L'équation d'Euler

Traduisons cette condition dans le cas de la fonctionnelle

$$J(y)=\int_a^b F\big(x,y(x),y'(x)\big)\,dx$$

où la fonction $F(x,y,p)$ a des dérivées partielles secondes continues. Rappelons que les fonctions admissibles y sont les fonctions deux fois dérivables telles que $y(a)=y_A$ et $y(b)=y_B$ sont des valeurs données.

Théorème. *Pour que $J(y)$ ait un extremum en $y=f$, il faut que*

$$\frac{\partial F}{\partial y}\big(x,f(x),f'(x)\big)-\frac{\partial}{\partial x}\left[\frac{\partial F}{\partial p}\big(x,f(x),f'(x)\big)\right]=0$$

L'égalité ci-dessus s'appelle *l'équation d'Euler*.

Démonstration. Pour un accroissement h tel que $h(a)=h(b)=0$, on a

$$J(y+h)-J(y)=\int_a^b\big[F(x,y+h,y'+h')-F(x,y,y')\big]\,dx$$

Écrivons l'approximation affine de $F(x,y,p)$ au point (x,y,y') :

$$F(x,y+h,y'+h')-F(x,y,y')=h\frac{\partial F}{\partial y}(x,y,y')+h'\frac{\partial F}{\partial p}(x,y,y')+\varphi(h)$$

où d'après le choix de la norme de h, $\varphi(h)$ est négligeable devant $\|h\|$. On en déduit que la variation de J en y est

$$L_y(h)=\int_a^b\left[h\frac{\partial F}{\partial y}\big(x,y(x),y'(x)\big)+h'\frac{\partial F}{\partial p}\big(x,y(x),y'(x)\big)\right]dx$$

Si $J(y)$ a un extremum local en f, alors $L_f(h)=0$ pour tout accroissement h deux fois dérivable tel que $h(a)=h(b)=0$. Posons pour simplifier $\alpha(x)=\dfrac{\partial F}{\partial y}\big(x,f(x),f'(x)\big)$ et $\beta(x)=\dfrac{\partial F}{\partial p}\big(x,f(x),f'(x)\big)$, de sorte que

(1) $$\int_a^b\big[\alpha(x)h(x)+\beta(x)h'(x)\big]\,dx=0\ ,\ \text{si } h(a)=h(b)=0.$$

► En posant $A(x)=\int_a^x\alpha(t)\,dt$ et en intégrant par parties, il vient

$$\int_a^b\alpha(x)h(x)\,dx=-\int_a^b A(x)h'(x)\,dx\,,$$

car $h(a)=h(b)=0$. La condition (1) devient

(2) $$\int_a^b\big[-A(x)+\beta(x)\big]h'(x)\,dx=\int_a^b u(x)h'(x)\,dx=0\ ,\ \text{où } u(x)=-A(x)+\beta(x).$$

► Posons $c=\dfrac{1}{b-a}\displaystyle\int_a^b u(x)\,dx$ et choisissons l'accroissement $h(x)=\displaystyle\int_a^x \big[u(t)-c\big]\,dt$. On a bien $h(a)=0$ et $h(b)=\displaystyle\int_a^b \big[u(t)-c\big]\,dt=\int_a^b u(t)\,dt-c(b-a)=0$. De plus,

$$\int_a^b \big[u(x)-c\big]h'(x)\,dx=\int_a^b u(x)h'(x)\,dx-c\big[h(b)-h(a)\big]=0,\quad \text{d'après (2).}$$

Or $h'(x)=u(x)-c$, donc il vient $\displaystyle\int_a^b \big[u(x)-c\big]^2\,dx=0$. Puisque la fonction $\big[u(x)-c\big]^2$ est continue et à valeurs positives ou nulles, on en déduit qu'elle est nulle (page 288). On a donc $u(x)=c$ pour tout $x\in[a,b]$.

Ainsi, nous avons montré que $\beta(x)-A(x)$ est constante, d'où $\beta'(x)=A'(x)=\alpha(x)$ pour tout x entre a et b : c'est l'équation d'Euler. ■

Exemple. Reprenons l'exemple précédent d'un rayon lumineux se propageant du point $A=(a,y_A)$ au point $B=(b,y_B)$ dans un milieu d'indice variable $n(x)$. Pour trouver la courbe $y(x)$ décrite par ce rayon, on doit minimiser la fonctionnelle $J(y)=\displaystyle\int_a^b n(x)\sqrt{1+y'(x)^2}\,dx$. Posons donc $F(x,y,p)=n(x)\sqrt{1+p^2}$. L'équation d'Euler s'écrit

$$0=\frac{\partial F}{\partial y}+\frac{\partial}{\partial x}\frac{\partial F}{\partial p}=\frac{\partial}{\partial x}\left(\frac{p\,n(x)}{\sqrt{1+p^2}}\right)\ ,\quad \text{car}\ \frac{\partial F}{\partial y}=0$$

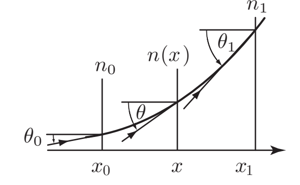
$$\frac{y'(x)\,n(x)}{\sqrt{1+y'(x)^2}}=K\ ,\quad \text{où } K \text{ est une constante.}$$

Il vient $y'=\dfrac{K}{\sqrt{n(x)^2-K^2}}$, ce qui ramène le calcul de $y(x)$ à celui d'une primitive. La solution générale dépend de deux constantes, K et la constante d'intégration dans la primitive : on les détermine au moyen des conditions $y(a)=y_A$ et $y(b)=y_B$.

En introduisant l'abscisse curviligne $ds=\sqrt{1+y'^2}\,dx$ (page 313), on a $n(x)\dfrac{dy}{dx}=K\dfrac{ds}{dx}$, donc $n(x)dy=K\,ds$ et $n(x)\dfrac{dy}{ds}=K$. Notons $\vec{T}$ le vecteur tangent unitaire à la courbe $y(x)$ et θ l'angle $\widehat{Ox,\vec{T}}$. Puisque $\dfrac{dy}{ds}=\sin\theta$, il vient

$$n(x)\sin\big(\theta(x)\big)=K$$

On obtient ainsi la « loi de la réfraction » : considérons une droite d'équation $x=x_0$; l'indice y est constant ; puisque la normale à la droite est dans la direction de l'axe des x, au point où il traverse cette droite, le rayon fait l'angle $\theta_0=\theta(x_0)$ avec la normale ; la loi de la réfraction affirme que pour deux points d'abscisses x_0 et x_1, on a $n_0\sin\theta_0=n_1\sin\theta_1$, où n_0 et n_1 sont les indices et θ_0, θ_1 les angles que fait le rayon avec la normale aux lignes (ou plus généralement aux surfaces) d'indice constant.



Cas particulier où F ne dépend pas de x

Supposons que la fonctionnelle est $J(y) = \int_a^b F(y, y')\, dx$. On a alors

$$\frac{d}{dx}\left[F - y'\frac{\partial F}{\partial p}\right] = \left[\frac{\partial F}{\partial y}y' + \frac{\partial F}{\partial p}y''\right] - y''\frac{\partial F}{\partial p} - y'\left[\frac{\partial^2 F}{\partial p^2}y'' + \frac{\partial^2 F}{\partial y \partial p}y'\right]$$

ou encore

$$(*) \qquad \frac{d}{dx}\left[F - y'\frac{\partial F}{\partial p}\right] = \frac{\partial F}{\partial y}y' - \frac{\partial^2 F}{\partial y \partial p}y'^2 - \frac{\partial^2 F}{\partial p^2}y''y'\,.$$

Dans l'équation d'Euler, l'expression est

$$\frac{\partial F}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial p}\right) = \frac{\partial F}{\partial y} - \frac{\partial}{\partial p}\left(\frac{\partial F}{\partial p}\right)\frac{dp}{dx} + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial p}\right)\frac{dy}{dx} = \frac{\partial F}{\partial y} - \frac{\partial^2 F}{\partial y \partial p}y' - \frac{\partial^2 F}{\partial p^2}y''$$

et en multipliant par y', on obtient $(*)$.

L'équation d'Euler s'écrit donc $\frac{d}{dx}\left[F - y'\frac{\partial F}{\partial p}\right] = 0$ et les solutions sont les fonctions f telles que

$$F\big(f(x), f'(x)\big) - f'(x)\frac{\partial F}{\partial p}\big(f(x), f'(x)\big) = c\ , \quad \text{où } c \text{ est une constante.}$$

4.2 Application : une stratégie de croissance végétale

Une plante partage ses ressources énergétiques entre croissance, reproduction et défense. Pour simplifier, considérons seulement le cas d'une plante qui, à tout âge t, consacre une partie $c(t)$ de ses ressources à sa croissance et le reste à la reproduction ; on a bien sûr $0 \leqslant c(t) \leqslant 1$ pour tout $t \geqslant 0$.

La *valeur reproductive* de la plante est

$$(1) \qquad R = \int_0^\infty q(t)\,\varphi(t)\,dt\,, \quad \text{où}$$

- $\varphi(t)$ est la quantité de graines produite à l'âge t,
- $q(t)$ est la probabilité que la plante atteigne l'âge t.

Voici une modélisation qui se propose de calculer la fonction $c(t)$.

Le facteur fécondité. Dans une espèce végétale, la fécondité dépend de l'âge : pour des individus pouvant atteindre une taille suffisante, la production moyenne de graines croît structurellement avec la taille (comme dans le cas d'un arbre, par exemple). Si l'on note $f(t)$ la fécondité, nous écrivons donc cette production sous la forme $\alpha\big(f(t)\big)^\beta$, où α et β sont des constantes allométriques positives.

Variation de la fécondité. Le taux de fécondité de la plante est le nombre dérivé $f'(t)$: faisons l'hypothèse qu'il est proportionnel à la croissance et à l'écart $f_m - f(t)$, où f_m est la fécondité correspondant à la taille maximum.

On a donc

$$f'(t) = a\big[f_m - f(t)\big]c(t)\ , \quad \text{où } a \text{ est un facteur positif propre à l'espèce.}$$

C'est une équation différentielle linéaire. La constante f_m est solution particulière et puisqu'on doit avoir $f(0) = 0$, il vient

$$(2) \qquad f(t) = f_m\big[1 - e^{-aC(t)}\big]\ , \quad \text{avec } C(t) = \int_0^t c(s)\,ds.$$

Évaluation de la production de graines. Considérons que pour un individu donné, la quantité de graines produite est proportionnelle à la part de ressources consacrée à sa reproduction. Puisque la plante ne consacre à sa reproduction qu'une partie $1 - c(t)$ de ses ressources, adoptons pour la fonction $\varphi(t)$ la formule

$$(3) \qquad \varphi(t) = \big[1 - c(t)\big]\alpha\big(f(t)\big)^{\beta}$$

Introduisons le taux de mortalité $\mu(t)$ à l'âge t et la probabilité $q(t)$ pour que la plante atteigne l'âge t. Entre des instants rapprochés t et $t+\delta t$, les plantes meurent en proportion $q(t)-q(t+\delta t)=\mu(t)q(t)\delta t$, d'où l'équation différentielle $q'(t)=-\mu(t)q(t)$. Puisque $q(0) = 1$, la solution est

$$(4) \qquad q(t) = e^{-M(t)}\ , \quad \text{avec } M(t) = \int_0^t \mu(s)\,ds.$$

En tenant compte des égalités (2), (3) et (4), la valeur reproductive (1) est ainsi

$$R = \alpha(f_m)^{\beta}\int_0^{\infty} e^{-M(t)}\big[1 - e^{-aC(t)}\big]^{\beta}\big[1 - c(t)\big]\,dt$$

On considère que la plante répartit ses ressources de manière à rendre maximum la valeur R. Puisque $c(t) = C'(t)$, introduisons la fonctionnelle

$$J(C) = \int_0^{\infty} e^{-M(t)}\big[1 - e^{-aC(t)}\big]^{\beta}\big[1 - C'(t)\big]\,dt$$

Pour découvrir la stratégie employée par la plante, il s'agit de trouver une fonction $C(t)$, définie pour $t\geqslant 0$, à dérivée $c(t)=C'(t)$ comprise entre 0 et 1 et telle que $J(C)$ est maximum. Bien qu'on ne soit pas exactement dans le cadre théorique précédent (car l'intégrale est généralisée), écrivons néanmoins l'équation d'Euler, en prenant la fonction

$$F(t, C, p) = e^{-M(t)}\big[1 - e^{-aC}\big]^{\beta}(1 - p)\,,$$

où p représente la variable $c = C'$. L'équation d'Euler est $\dfrac{\partial F}{\partial C} - \dfrac{\partial}{\partial t}\dfrac{\partial F}{\partial p} = 0$, avec

- $\dfrac{\partial F}{\partial C} = e^{-M(t)}\beta\big[1 - e^{-aC}\big]^{\beta-1}ae^{-aC}(1 - c)$
- $\dfrac{\partial F}{\partial p} = -e^{-M(t)}\big[1 - e^{-aC}\big]^{\beta}$
- $\dfrac{\partial}{\partial t}\dfrac{\partial F}{\partial p} = \mu(t)e^{-M(t)}\big[1 - e^{-aC}\big]^{\beta} - \beta e^{-M(t)}\big[1 - e^{-aC}\big]^{\beta-1}ace^{-aC}$

et il vient après simplification

$$0 = e^{-M(t)}\beta ae^{-aC(t)}\big[1 - e^{-aC(t)}\big]^{\beta-1} - \mu(t)e^{-M(t)}\big[1 - e^{-aC(t)}\big]^{\beta}$$

On en déduit $\beta a e^{-aC(t)} - \mu(t)\left[1 - e^{-aC(t)}\right] = 0$, ou encore $e^{aC(t)} = \dfrac{\mu(t) + \beta a}{\mu(t)}$ et finalement

(5) $$C(t) = \frac{1}{a} \ln \left[1 + \frac{\beta a}{\mu(t)}\right]$$

La fonction $C(t)$ est la quantité totale de ressources que la plante a affectée à sa croissance jusqu'à l'âge t.
On considère habituellement qu'avant d'atteindre un âge t_0 de maturité reproductive, la plante consacre toutes ses ressources à sa croissance afin d'atteindre au plus vite sa période de reproduction : on a donc $c(t) = 1$ pour $t \leqslant t_0$ et par suite $C(t) = t$ pour $t \leqslant t_0$. Ainsi, la solution $C(t)$ est en réalité définie en deux morceaux :

$$C(t) = \begin{cases} t & \text{si } 0 \leqslant t \leqslant t_0 \\ \dfrac{1}{a} \ln \left[1 + \dfrac{\beta a}{\mu(t)}\right] & \text{si } t \geqslant t_0 \end{cases}$$

L'instant t_0 est l'âge de première reproduction : il satisfait l'équation $t_0 = \dfrac{1}{a} \ln \left[1 + \dfrac{\beta a}{\mu(t_0)}\right]$, car C est une fonction continue. La figure 1 montre l'allure d'une courbe $C(t)$ typique. Sur la figure 2, on voit la dérivée $c(t) = C'(t)$ sur chaque intervalle : c'est la proportion des ressources que la plante affecte à chaque instant à sa croissance.

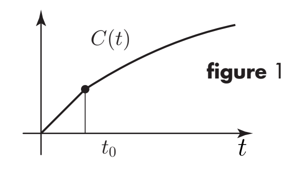


figure 1

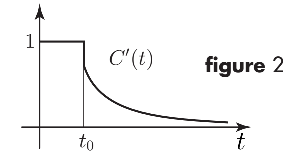
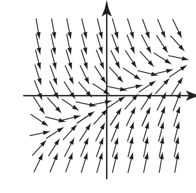


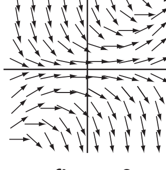
figure 2

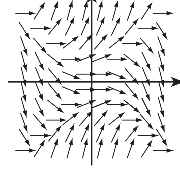
Exercices

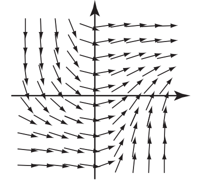
@ **1.** Associer chaque équation différentielle à son champ de directions.

a) $x' = x^2 - t^2$ **b)** $x' = t - 2x$ **c)** $x' = te^{-tx}$ **d)** $x' = x(t - x)$


figure 1


figure 2


figure 3


figure 4

@ **2. Forme normale d'une équation linéaire du second ordre.** Soit l'équation différentielle (1) $x'' + p(t)x' + q(t)x = 0$, où $p(t)$ et $q(t)$ sont des fonctions.

a) Montrer que par le changement d'inconnue $x(t) = z(t)a(t)$, l'équation (1) devient $az'' + (2a' + pa)z' + (a'' + pa' + qa)z = 0$.

b) Posons $a(t) = \exp\left[\frac{-1}{2} \int_{t_0}^{t} p(s)\, ds\right]$. Montrer que l'on a $a(t) > 0$ pour tout t et que

z est solution de l'équation différentielle linéaire (2) $z'' + \big(q - \frac{p'}{2} - \frac{p^2}{4}\big)z = 0$ où la dérivée première de z ne figure plus.

c) Utiliser cette transformation pour résoudre l'équation $x'' + 2tx' + t^2x = 0$ (on se ramène à $z'' - z = 0$).

@ 3. Le wronskien. Soit l'équation différentielle (1) $x'' + p(t)x' + q(t)x = 0$, où $p(t)$ et $q(t)$ sont des fonctions. Si $x(t)$ et $y(t)$ sont des solutions de (1), leur *wronskien* est la fonction $w = xy' - x'y$.

a) Montrer que $w' = xy'' - x''y$ et que $w' + pw = 0$. En déduire que $w(t) = w(t_0)\exp\big[-\int_{t_0}^{t} p(s)\,ds\big]$.

On peut ainsi, directement à partir de l'équation différentielle (1), calculer le wronskien de deux solutions à un facteur multiplicatif près.

b) Supposons que I est un intervalle où $x(t)$ ne s'annule pas. Montrer que $\left(\frac{y}{x}\right)' = \frac{w}{x^2}$ et que pour t_0 et t dans I, on a $y(t) = x(t)\int^{t} \frac{w(s)}{x(s)^2}\,ds$.

Si l'on connaît une solution x, cette formule fournit une solution y non proportionnelle à x.

c) Appliquons ce qui précède à l'équation de Bessel (b_0) $x'' + \frac{1}{t}x' + x = 0$ dont une solution est la fonction $J_0(t)$ définie dans l'exercice **8** page 394. Montrer que si α et β sont deux zéros consécutifs et de même signe de J_0, alors entre α et β les solutions de (b_0) sont les fonctions $cJ_0(t) + dJ_0(t)\int^{t} \frac{ds}{sJ_0(s)^2}$, où c et d sont des constantes.

@ 4. Un exemple d'utilisation du wronskien. Supposons que $x(t)$ est une solution de l'équation de Bessel (b_0) : $tx'' + x' + tx = 0$ et que $x(t)$ est définie en $t = 0$.

a) Montrer que le wronskien $w = x'J_0 - xJ_0'$ est défini au voisinage de $t = 0$ et solution de l'équation $tw' + w = 0$. Résoudre cette équation et en déduire que l'on a $w(t) = 0$ pour tout t.

b) Montrer que $\big(\frac{x}{J_0}\big)' = 0$. En déduire que l'on a $x(t) = x(0)J_0(t)$ quel que soit t.

Pour l'équation (b_0), les seules solutions définies en $t = 0$ sont donc les fonctions cJ_0, où c est une constante.

@ 5. Équations oscillantes. On considère l'équation différentielle $x'' + q(t)x = 0$ et l'on suppose qu'il existe un nombre $\omega > 0$ tel que, pour tout t, $q(t) \geqslant \omega^2$.
Soit t_0 un nombre réel ; on pose $t_1 = t_0 + \frac{\pi}{\omega}$ et $S(t) = \sin\big[\omega(t - t_0)\big]$.
Nous allons voir que si $q(t)$ est définie sur l'intervalle $[t_0, t_1]$, alors toutes les solutions $x(t)$ de l'équation différentielle $x'' + q(t)x = 0$ s'annulent au moins une fois dans l'intervalle $[t_0, t_1]$.

a) Posons $w = x'S - xS'$. Montrer que $w'(t) = S(t)x(t)\big[\omega^2 - q(t)\big]$.

b) Supposons $x(t) > 0$ pour $t_0 \leqslant t \leqslant t_1$. Montrer qu'alors on a $w(t_0) < 0$, $w(t_1) > 0$ et que w est décroissante sur $[t_0, t_1]$. En déduire une contradiction et le résultat annoncé.

c) Supposons que l'on a $q(t) \geqslant \omega^2$ sur un intervalle I de longueur au moins π/ω. Montrer qu'alors chaque solution de (1) s'annule au moins une fois dans tout intervalle J inclus dans I et de longueur π/ω. Si par exemple $I =]-\infty, +\infty[$ ou bien $I =]a, +\infty[$, alors chaque solution a une infinité de zéros dans I : on dit que l'équation différentielle (1) est oscillante.

L'équation $x'' + \omega^2 x = 0$ a pour solutions les fonctions $\sin(\omega t + \varphi)$ qui oscillent d'autant plus vite que ω est grand. Le résultat précédent montre que dans l'équation (1), plus $q(t)$ est grand, plus les solutions oscillent vite.

d) On a esquissé page 158 le graphe d'une solution de l'équation d'Airy $x''+tx=0$: que peut-on dire de la courbe dans le domaine $t > 0$? (voir aussi l'exercice **7** page 572)

6. L'équation de Bessel générale. Étant donné un nombre réel α, l'équation de Bessel d'indice α est (b_α) : $x'' + \frac{1}{t}x' + \left(1 - \frac{\alpha^2}{t^2}\right)x = 0$, définie dans l'intervalle $]0, +\infty[$. Posons $x(t) = y(t)/\sqrt{t}$.

a) Montrer que $y(t)$ est solution de l'équation $y'' + \left(1 - \frac{\alpha^2 - (1/4)}{t^2}\right)y$.

b) En appliquant l'exercice précédent, montrer que toutes les solutions de (b_α) sont oscillantes dans $]0, +\infty[$ et que si $|\alpha| \leqslant 1/2$, alors chaque solution de (b_α) s'annule au moins une fois dans tout intervalle de longueur π.

Le graphe de la solution J_0 de (b_0) est montré page 394. Les valeurs des premiers zéros positifs des fonctions J_0 et $J_1 = -J_0'$ sont données page 577.

c) Résoudre l'équation de Bessel pour $\alpha = 1/2$.

L'équation de Bessel intervient dans de nombreuses questions de Physique, notamment dans l'étude des vibrations d'une membrane de tambour. En appelant r, θ les coordonnées polaires dans le plan de la membrane et $z(r, \theta, t)$ l'élévation à l'instant t du point de coordonnées (r, θ), les solutions fondamentales sont de la forme $z = \sin(\omega t)\sin(n\theta)u(r)$, où n est un entier positif. La fonction $u(r)$ doit vérifier l'équation $u'' + \frac{1}{r}u' + \left(\omega^2 c^2 - \frac{n^2}{r^2}\right)u = 0$, où c est un coefficient qui dépend des unités, de la tension de la membrane et de son élasticité. Il est facile de voir qu'en posant $v(\omega c r) = u(r)$, la fonction v est solution de l'équation de Bessel (b_n). Il faut que u soit définie en $r = 0$, ce qui impose $u(0) = 0$, et si a est le rayon de la membrane, on doit avoir $u(a) = 0$: il s'ensuit que pour n fixé, ω doit être l'un des nombres $z_1/ca, \ldots, z_k/ca, \ldots$, où $z_1, \ldots, z_k \ldots$ sont les zéros de v.

7. Équation linéaire du second ordre où l'on connaît une solution de l'équation homogène
Soit l'équation différentielle (1) $x'' + p(t)x' + q(t)x = b(t)$, où $p(t)$, $q(t)$ et $b(t)$ sont des fonctions. Supposons que $u(t)$ est une solution de l'équation homogène (h) $x'' + p(t)x' + q(t)x = 0$ et cherchons une fonction $\lambda(t)$ non constante telle que $v(t) = \lambda(t)u(t)$ soit solution de (1).
Montrer que $v'' + pv + qv' = u\lambda'' + (pu + 2u')\lambda' + (u'' + pu' + qu)\lambda$. En déduire que $v(t)$ est solution si et seulement si la fonction $z = \lambda'$ est solution de l'équation du

premier ordre $uz' + (pu + 2u')z = b(t)$. On sait que cette dernière équation se résout au moyen de primitives.

8. Longues vagues dans un estuaire. Considérons un estuaire de longueur a s'ouvrant sur la mer. Prenons un axe Ox comme sur la figure ci-contre, l'origine étant au fond de l'estuaire. La largeur b de l'estuaire et sa profondeur h sont des fonctions de x. Par rapport au niveau de la surface au repos, l'élévation de l'eau dans l'estuaire, à l'abscisse x et à l'instant t, est une fonction $y(x,t)$ solution de l'équation $\dfrac{\partial^2 y}{\partial t^2} = \dfrac{g}{b}\dfrac{\partial}{\partial x}\left(hb\dfrac{\partial y}{\partial x}\right)$ (à condition de négliger les écarts de pression dûs aux accélérations verticales des particules liquides).
À l'embouchure ($x = a$), il se maintient une houle périodique $C\cos(\omega t + \varphi)$, de sorte que l'équation devient

$$\frac{g}{b}\frac{\partial}{\partial x}\left(hb\frac{\partial y}{\partial x}\right) + \omega^2 y = 0\,.$$

Supposons la profondeur h constante et la largeur b proportionnelle à x.

a) Montrer que $y(x,t) = z(x)\cos(\omega t + \varphi)$, où z est solution de l'équation $z'' + \dfrac{1}{x}z' + k^2 z$, avec $k = \omega/\sqrt{gh}$.

b) En utilisant l'exercice **4**, montrer que $y(x,t) = C\dfrac{J_0(kx)}{J_0(ka)}\cos(\omega t + \varphi)$.

c) En se reportant au graphique de l'exercice **8** page 394, expliquer pourquoi dans l'estuaire, l'amplitude des vagues augmente au fur et à mesure qu'elles remontent depuis l'embouchure, alors que la distance entre deux crêtes successives reste presque constante.

@ **9. Expansion d'une bulle de gaz.** Une bulle de gaz sphérique immergée dans un fluide subit une expansion sous l'effet de sa pression interne p. Soit r_0 le rayon initial et p_i la pression interne à l'instant $t = 0$. Supposons que la pression ambiante est négligeable devant p et que l'expansion suit la loi adiabatique $\dfrac{p}{p_i} = \left(\dfrac{r_0}{r}\right)^{3\gamma}$, où γ est l'exposant caractéristique du gaz. Notons ρ la masse volumique du fluide et $r(t)$ le rayon de la bulle à l'instant t. En faisant l'hypothèse $\dot r = 0$ à $t = 0$, r satisfait l'équation

$$(1) \qquad r\ddot r + \frac{3}{2}\dot r^2 = c^2\left(\frac{r_0}{r}\right)^{3\gamma}$$

où $c^2 = p_i/\rho$ (le coefficient c est homogène à une vitesse).

a) En multipliant (1) par $r^2\dot r$, montrer que l'on a $\dfrac{d}{dt}\left(r^3\dot r^2\right) = \dfrac{d}{dt}\left(\dfrac{2c^2 r_0^{3\gamma}}{3-3\gamma}\,r^{3-3\gamma}\right)$.

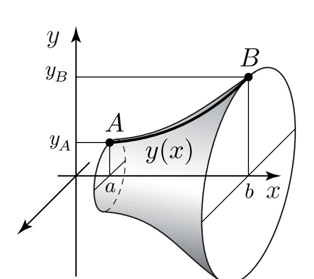
b) En déduire l'intégrale première $\dfrac{\dot r^2}{c^2} = \dfrac{2}{3(\gamma-1)}\left[\left(\dfrac{r_0}{r}\right)^3 - \left(\dfrac{r_0}{r}\right)^{3\gamma}\right]$.

c) Prenons $\gamma = 4/3$ (valeur convenable pour un gaz diatomique) et posons $r = r_0(1+z)$. Montrer que l'on a $(1+z)^2\dot z = \dfrac{c}{r_0}\sqrt{2z}$. En déduire $\sqrt{2z}\left(1 + \dfrac{2}{3}z + \dfrac{1}{5}z^2\right) = \dfrac{c}{r_0}t$.

d) Supposons par exemple que le diamètre initial de la bulle est 1 mètre et sa pression initiale de 10^3 atmosphères (une atmosphère vaut à peu près 10^3 hPa). Montrer que, dans l'eau, le rayon de la bulle double en $1/250$ seconde environ et est multiplié par 5 en $1/30$ seconde environ.

Ce modèle est utilisé aussi bien en Astrophysique que dans le cas de l'explosion d'une mine sous-marine ou dans les dispositifs de sécurité pour la soudure à l'arc en plongée.

10. Surface de révolution d'aire minimum. Parmi toutes les courbes $y(x)$ joignant deux points $A = (a, y_A)$ et (b, y_B), cherchons celle qui engendre une surface d'aire minimum quand on la fait tourner autour de l'axe Ox.



Pour une courbe située dans le demi-plan $y > 0$, l'aire en question est $J(y) = 2\pi \int_a^b y\sqrt{1+y'^2}\,dx$. Supposons que $y(x)$ est une solution du problème.

a) En remarquant que $F(y,p) = y\sqrt{1+p^2}$ ne dépend pas de x, montrer que l'équation d'Euler conduit à la relation $y\sqrt{1+y'^2} - y'\dfrac{yy'}{\sqrt{1+y'^2}} = c$, où c est une constante.

b) En déduire que l'on a $y(x) = c\sqrt{1+y'^2}$. En supposant $y_B > y_A$, montrer que $y' = \dfrac{1}{c}\sqrt{y^2-c^2}$ et $x + d = c\ln \dfrac{y + \sqrt{y^2-c^2}}{c}$.

c) Montrer que si l'on pose $u = \dfrac{1}{c}(y + \sqrt{y^2-c^2})$, alors $\dfrac{1}{u} = \dfrac{1}{c}(y - \sqrt{y^2-c^2})$. En déduire $y(x) = \dfrac{c}{2}\left(e^{(x+d)/c} + e^{-(x+d)/c}\right) = \dfrac{c}{2}\,\text{ch}\,\dfrac{x+d}{c}$, où ch est la fonction cosinus hyperbolique (page 268).

On doit choisir les constantes de manière que $y(a) = y_A$ et $y(b) = y_B$. Si ces conditions déterminent c et d de manière unique, alors le problème a une seule solution. S'il n'y a pas de constantes c et d possibles, il n'y a pas de solution.

@11. Une équation autonome. Soit l'équation différentielle $(*)$ $x' = x^2 - x^3$.

a) Quelles sont les solutions constantes ?

b) Quel est le sens de variation de la solution telle que $x(0) = 1/2$? Quel est son intervalle de définition ? Quelles sont ses limites aux bornes de cet intervalle ?

c) Soit s la solution de $(*)$ telle que $s(0) = 2$. Quelle est la relation entre $s(t)$ et t ? (trouver un polynôme U de degré 1 et un nombre p tels que $\dfrac{U}{x^2} + \dfrac{p}{1-x} = \dfrac{1}{x^2-x^3}$ et utiliser la méthode page 323 : la relation est $\ln\left[\dfrac{s(t)}{s(t)-1}\right] - \dfrac{1}{s(t)} = t + \ln 2 - \dfrac{1}{2}$)

d) Montrer que l'intervalle de définition de s est de la forme $]a, +\infty[$ et calculer le nombre a. Montrer que s est décroissante et que $\lim_{t\to+\infty} s(t) = 1$.

e) Montrer que $s(t) \sim \dfrac{1}{\sqrt{2(t-a)}}$ quand t tend vers a (utiliser **(c)**).

Chapitre 16

Systèmes différentiels

Quand les variations d'une quantité x dépendent de sa valeur, son comportement est régi par une équation différentielle $x' = f(t, x)$. Lorsque plusieurs quantités x, y sont en jeu et que chacune influe aussi sur le taux de variation de l'autre, les équations sont de la forme $\begin{cases} x' = f(t,x,y) \\ y' = g(t,x,y) \end{cases}$, où f et g sont des fonctions : c'est un *système différentiel.* Une solution est constituée de deux fonctions $x(t)$, $y(t)$ vérifiant $x'(t) = f\big(t, x(t), y(t)\big)$ et $y'(t) = g\big(t, x(t), y(t)\big)$ pour tout t.

Exemple : proies et prédateurs. En l'absence de prédateurs, une espèce animale (A) se reproduit a un taux proportionnel au nombre d'individus (loi de Malthus) : cela veut dire que si x est le nombre d'individus à un certain instant t, la vitesse $\dot{x}$ d'accroissement de population est ax, où a est une constante positive. La fonction $x(t)$ est solution de l'équation différentielle $\dot{x}{=}ax$: si x_0 est le nombre d'individus à l'instant $t = 0$, la solution $x_0 e^{at}$ croît exponentiellement vers l'infini, ce qui n'est pas réaliste. Supposons que l'espèce (A) est soumise à des prédateurs (P) en nombre $y(t)$ et faisons des hypothèses vraisemblables sur le comportement de ces populations.

i) La mortalité chez (A) due à la prédation est de la forme bxy, proportionnelle au nombre xy de contacts entre les individus.

ii) Le taux de croissance des prédateurs est proportionnel à leur nombre y et à la quantité de proies, donc de la forme dxy, où d est une constante.

iii) La surpopulation des prédateurs est évitée grâce à un taux de mortalité cy proportionnel à leur nombre (autolimitation due à la compétition pour la nourriture).

Cela conduit au système différentiel suivant :

$$\begin{cases} \dot{x} = ax - bxy \\ \dot{y} = -cy + dxy \end{cases}, \quad \text{où les constantes } a, b, c, d \text{ sont positives.}$$

Système différentiel autonome. Dans l'exemple précédent, le système différentiel s'écrit $\begin{cases} x' = f(x,y) \\ y' = g(x,y) \end{cases}$, où les expressions $f(x, y)$ et $g(x, y)$ ne dépendent pas

de t : on dit que le système est autonome. Plus généralement, un système différentiel est *autonome* si la relation entre fonctions et dérivées y est indépendante du temps.

Comment rendre autonome un système différentiel
Considérons par exemple le système non autonome

$$(1)\qquad \begin{cases} x' &= f(t,x,y) \\ y' &= g(t,x,y) \end{cases}$$

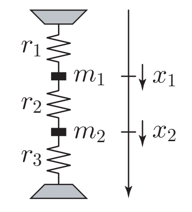
En plus des variables d'état x et y, introduisons une nouvelle variable d'état τ et considérons le système différentiel

$$(2)\qquad \begin{cases} \tau' &= 1 \\ x' &= f(\tau,x,y) \\ y' &= g(\tau,x,y) \end{cases}$$

C'est un système autonome à trois fonctions inconnues $\tau(t)$, $x(t)$ et $y(t)$. La première équation donne $\tau(t) = t + c$, où c est une constante. Si l'on résout (2) avec la condition initiale $\tau(t_0) = t_0$, alors on a $\tau(t) = t$ quel que soit t et les fonctions $x(t)$ et $y(t)$ sont solutions de (1). Tout système différentiel peut de cette façon se ramener à un système autonome de taille un de plus.

1. Systèmes différentiels linéaires

Exemple 1. Dans le dispositif représenté ci-contre, les masses m_1, m_2 sont fixées à des supports par l'intermédiaire de ressorts r_1, r_3 et reliées entre-elles par le ressort r_2 ; l'ensemble est mobile verticalement. Repérons la position des masses le long d'un axe vertical en appelant x_1, x_2 l'écart de chacune par rapport à sa position d'équilibre. Les équations différentielles du mouvement sont



$$(1)\qquad \begin{cases} m_1\ddot{x}_1 &= -k_1x_1 + k_2(x_2-x_1) \\ m_2\ddot{x}_2 &= -k_2(x_2-x_1) - k_3x_2 \end{cases}$$

où k_1, k_2, k_3 sont les raideurs des ressorts. En posant $v_1 = \dot{x}_1$ et $v_2 = \dot{x}_2$, ce système devient

$$\begin{cases} \dot{x}_1 &= v_1 \\ \dot{x}_2 &= v_2 \\ \dot{v}_1 &= -\big[(k_1+k_2)/m_1\big]x_1 + (k_2/m_1)x_2 \\ \dot{v}_2 &= (k_2/m_2)x_1 - \big[(k_2+k_3)/m_2\big]x_2 \end{cases}$$

Une solution $\big(x_1(t), x_2(t), v_1(t), v_2(t)\big)$ est représenté par un point mobile dans l'espace des quatre coordonnées (x_1, x_2, v_1, v_2) ; cet espace, appelé *l'espace des phases* du système, est de dimension 4. Une solution définit une courbe paramétrée, donc une trajectoire, dans l'espace des phases. Introduisons le vecteur $X(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ v_1(t) \\ v_2(t) \end{bmatrix}$ et rap-

pelons que le vecteur dérivé $\dot{X}(t)$ s'obtient en dérivant coordonnée par coordonnée. En utilisant la notation matricielle, le système différentiel ci-dessus s'écrit

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{v}_1(t) \\ \dot{v}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -(k_1+k_2)/m_1 & k_2/m_1 & 0 & 0 \\ k_2/m_2 & -(k_2+k_3)/m_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ v_1(t) \\ v_2(t) \end{bmatrix} \quad \text{ou encore}$$

$$(S) \qquad \dot{X} = AX \quad , \quad \text{avec } A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -(k_1+k_2)/m_1 & k_2/m_1 & 0 & 0 \\ k_2/m_2 & -(k_2+k_3)/m_2 & 0 & 0 \end{bmatrix}$$

La matrice A s'appelle la matrice du système. Le système différentiel (1) est d'ordre 2, car il fait intervenir des dérivées secondes, mais le système différentiel (S), de taille 4, ne comporte plus que des dérivées premières.

Exemple 2. En Économie, on distingue deux types de taux d'intérêt : un taux à court terme c et un taux à long terme ℓ. Tout changement de valeur pour l'un des taux a une répercussion sur l'autre. Voici un modèle simple pour les variations conjointes de c et ℓ au cours du temps :

$$\begin{cases} \dot{c} = a(r - c - \ell) \\ \dot{\ell} = b(r - c - \ell) \end{cases} \quad \text{ou encore} \quad \begin{bmatrix} \dot{c} \\ \dot{\ell} \end{bmatrix} = \begin{bmatrix} -a & -a \\ -b & -b \end{bmatrix} \begin{bmatrix} c \\ \ell \end{bmatrix} + \begin{bmatrix} ar \\ br \end{bmatrix}$$

Les constantes r, a, b sont positives : a et b sont des vitesses d'évolution pour les taux c et ℓ ; r s'interprète comme la valeur limite de $c + \ell$ (exercice 1). Introduisons le vecteur d'état $X(t) = \begin{bmatrix} c(t) \\ \ell(t) \end{bmatrix}$, la matrice $A = \begin{bmatrix} -a & -a \\ -b & -b \end{bmatrix}$ du système et le vecteur constant $B = \begin{bmatrix} ar \\ br \end{bmatrix}$. Le système s'écrit $\dot{X} = AX + B$ et le vecteur B s'appelle le second membre du système.

Exemple 3. Le butène, un corps chimique de formule C_4H_8, se présente sous trois formes isomères B_1, B_2, B_3, correspondant à différentes dispositions des atomes dans la molécule.

butène-1: $CH_2 = CH - CH_2 - CH_3$

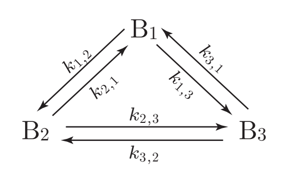
cis-butène-2 / trans-butène-2

cis-butène-2 : CH_3 (haut gauche), CH_3 (haut droite), $C = C$, H (bas gauche), H (bas droite)

trans-butène-2 : H (haut gauche), CH_3 (haut droite), $C = C$, CH_3 (bas gauche), H (bas droite)

Ces trois formes coexistent et s'échangent au cours de réactions équilibrées, les concentrations $[\mathrm{B}_1]$, $[\mathrm{B}_2]$, $[\mathrm{B}_3]$ variant au cours du temps selon les équations suivantes :

$$\begin{cases} \dfrac{d[\mathrm{B}_1]}{dt} = -(k_{1,2}+k_{1,3})[\mathrm{B}_1] + k_{2,1}[\mathrm{B}_2] + k_{3,1}[\mathrm{B}_3] \\ \dfrac{d[\mathrm{B}_2]}{dt} = k_{1,2}[\mathrm{B}_1] - (k_{2,1}+k_{2,3})[\mathrm{B}_2] + k_{3,2}[\mathrm{B}_3] \\ \dfrac{d[\mathrm{B}_3]}{dt} = k_{1,3}[\mathrm{B}_1] + k_{2,3}[\mathrm{B}_2] - (k_{3,1}+k_{3,2})[\mathrm{B}_3] \end{cases}$$



où les nombres positifs $k_{i,j}$ sont constants à température et pression données. En introduisant le vecteur d'état $X = \begin{bmatrix} [\mathrm{B}_1] \\ [\mathrm{B}_2] \\ [\mathrm{B}_3] \end{bmatrix}$ et la matrice du système, les équations s'écrivent

$$\frac{dX}{dt} = AX \quad , \quad \text{où} \quad A = \begin{bmatrix} -k_{1,2}-k_{1,3} & k_{2,1} & k_{3,1} \\ k_{1,2} & -k_{2,1}-k_{2,3} & k_{3,2} \\ k_{1,3} & k_{2,3} & -k_{3,1}-k_{3,2} \end{bmatrix}.$$

Les proportions dans le mélange évoluent vers un équilibre caractérisé par $\dfrac{d[\mathrm{B}_1]}{dt} = \dfrac{d[\mathrm{B}_2]}{dt} = \dfrac{d[\mathrm{B}_3]}{dt} = 0$. Remarquons que la somme $[\mathrm{B}_1] + [\mathrm{B}_2] + [\mathrm{B}_3]$ des concentrations est constante, car en ajoutant les lignes du système, il vient $\dfrac{d[\mathrm{B}_1]}{dt} + \dfrac{d[\mathrm{B}_2]}{dt} + \dfrac{d[\mathrm{B}_3]}{dt} = 0$.

Dans les exemples précédents, les systèmes différentiels introduits sont de la forme

$$\begin{cases} x_1' = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1(t) \\ x_2' = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2(t) \\ \vdots \qquad\qquad \vdots \\ x_n' = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_n(t) \end{cases}$$

où les coefficients a_{ij} sont des constantes et où $b_1(t), \ldots, b_n(t)$ sont des fonctions. En posant

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \text{et} \quad B(t) = \begin{bmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{bmatrix},$$

le système s'écrit $\quad X' = AX + B(t)$.

Définitions

Un système différentiel de la forme $X' = AX + B(t)$, où A est une matrice carrée à coefficients constants et où $B(t)$ est un vecteur de fonctions, s'appelle un *système différentiel linéaire à coefficients constants*. La matrice A est la *matrice du système* et la fonction $t \mapsto B(t)$ est le *second membre*. Si $B(t) = 0$ quel que soit t, le système est dit *homogène*. Le système différentiel $X' = AX$ s'appelle le système homogène associé.

Une *solution* du système différentiel $X' = AX + B(t)$ est une fonction $X(t)$ à valeurs vectorielles telle que $X'(t) = AX(t) + B(t)$ pour tout t.

- Supposons que $X_1(t)$ et $X_2(t)$ sont des solutions du système différentiel homogène $X' = AX$ et formons la combinaison linéaire $Y(t) = \alpha_1 X_1(t) + \alpha_2 X_2(t)$, où α_1, α_2 sont des nombres. Alors on a $X_1'(t) = AX_1(t)$, $X_2'(t) = AX_2(t)$ et en ajoutant il vient

$$Y' = \alpha_1 X_1' + \alpha_2 X_2' = \alpha_1 AX_1 + \alpha_2 AX_2 = A\big(\alpha_1 X_1 + \alpha_2 X_2\big) = AY$$

Pour un système différentiel linéaire homogène $X' = AX$, toute combinaison linéaire de solutions est encore solution.

- Supposons que $S(t)$ est une solution du système différentiel $X' = AX + B(t)$. Puisque $S'(t) = AS(t) + B(t)$, une fonction $X(t)$ est solution si et seulement si

$$X' - S' = \big(AX + B(t)\big) - \big(AS + B(t)\big) \iff (X - S)' = AX - AS = A(X - S)$$

Ainsi, pour que $X(t)$ soit solution de $X' = AX + B(t)$, il faut et il suffit que la fonction $X(t) - S(t)$ soit solution du système homogène associé.

Propriétés générales des solutions

Pour le système homogène $X' = AX$: *la fonction nulle est solution et toute combinaison linéaire de solutions est solution ; l'ensemble des solutions est un espace vectoriel.*

Pour le système complet $X' = AX + B(t)$: *si $S(t)$ est une solution de $X' = AX + B(t)$, alors toutes les solutions sont de la forme $X(t) = X_0(t) + S(t)$, où $X_0(t)$ est une solution quelconque du système homogène $X' = AX$.*
La différence de deux solutions est donc une solution du système homogène.

Pour résoudre le système différentiel, il faut en trouver une solution particulière et résoudre le système homogène associé.

1.1 Résolution du système homogène $X' = AX$

Exemple 1. Prenons $A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$. Le système différentiel $X' = AX$ s'écrit

$$\begin{cases} x_1' = ax_1 \\ x_2' = bx_2 \end{cases}$$

La première équation est une équation différentielle de la seule fonction inconnue x_1. Les solutions sont $x_1(t) = k_1 e^{at}$, où k_1 est une constante quelconque. De même, les solutions de la seconde équation sont les fonctions $x_2(t) = k_2 e^{bt}$, où k_2 est quelconque. Les solutions du système sont les fonctions

$$X(t) = \begin{bmatrix} k_1 e^{at} \\ k_2 e^{bt} \end{bmatrix} = \begin{bmatrix} e^{at} & 0 \\ 0 & e^{bt} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

où k_1 et k_2 sont des constantes. On a $X(0) = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$, c'est-à-dire $x_1(0) = k_1$ et $x_2(0) = k_2$: la solution est donc entièrement déterminée par sa valeur en $t = 0$.

Exemple 2. Posons $A = \begin{bmatrix} a & p \\ 0 & a \end{bmatrix}$. Le système différentiel $X' = AX$ s'écrit

$$\begin{cases} x_1' = ax_1 + px_2 \\ x_2' = ax_2 \end{cases}$$

Commençons par la seconde équation qui n'a qu'une fonction inconnue. Ses solutions sont $x_2(t) = k_2 e^{at}$, où k_2 est quelconque. En reportant dans la première équation, il vient

$$(*) \qquad x_1' = ax_1 + pk_2 e^{at}$$

une équation du premier ordre avec second membre. La solution générale de l'équation homogène est $k_1 e^{at}$. Cherchons une solution particulière de $(*)$ sous la forme $s(t) = c(t)e^{at}$, conformément à la méthode de variation de la constante (page 444). En reportant $s(t)$ dans $(*)$, il vient

$$ac(t)e^{at} + c'(t)e^{at} = ac(t)e^{at} + pk_2 e^{at}$$

et après simplification, on obtient $c'(t) = pk_2$, dont une solution est $c(t) = pk_2 t$. La fonction $s(t) = pk_2 te^{at}$ est donc une solution de $(*)$. On en déduit que les solutions de $(*)$ sont les fonctions

$$x_1(t) = k_1 e^{at} + k_2 pte^{at} \ , \quad \text{où } k_1 \text{ est une constante quelconque.}$$

Finalement, les solutions de $X' = AX$ sont les fonctions

$$X(t) = \begin{bmatrix} k_1 e^{at} + k_2 pte^{at} \\ k_2 e^{at} \end{bmatrix} = \begin{bmatrix} e^{at} & pte^{at} \\ 0 & e^{at} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

où k_1 et k_2 sont des constantes. On a $X(0) = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$, donc $X(t)$ est déterminée par sa valeur en $t = 0$.

Ces exemples montrent que si la matrice A est diagonale ou triangulaire, le système se résout en prenant les équations une à une : on commence par l'équation qui ne contient qu'une inconnue et l'on substitue les solutions dans les autres, de proche en proche. Pour résoudre un système homogène général, nous allons nous ramener au cas d'une matrice diagonale ou triangulaire.
Faisons un changement d'inconnue en posant $X = PY$, où P est une matrice carrée à coefficients constants. On suppose P inversible, de manière à pouvoir exprimer $Y = P^{-1}X$ au moyen de X.
Chaque coefficient $x_i(t)$ de $X(t)$ est une combinaison $p_1y_1(t) + p_2y_2(t) + \cdots + p_ny_n(t)$, où $[\, p_1 \ p_2 \ \cdots \ p_n \,]$ est la i-ème ligne de P. On a $x_i'(t) = p_1y_1'(t) + p_2y_2'(t) + \cdots + p_ny_n'(t)$, donc $X'(t) = PY'(t)$. Il vient

$$X' = AX \iff PY' = APY \iff Y' = P^{-1}APY$$

Pour que la fonction $X(t) = PY(t)$ soit solution de $X' = AX$, il faut et il suffit que $Y(t)$ soit solution du système différentiel $Y' = (P^{-1}AP)Y$.

Méthode de résolution

On utilise la technique de diagonalisation ou de trigonalisation expliquée pages 176 à 182. Distinguons plusieurs cas.

A est diagonalisable avec des valeurs propres réelles. On calcule les valeurs propres $\lambda_1, \dots, \lambda_n$ de A et les vecteurs propres associés $V_1, V_2, \dots, V_n$.

- Puisque A est diagonalisable, les vecteurs propres forment une base de l'espace vectoriel $\mathbb{R}^n$, la matrice P dont les colonnes sont $V_1, \dots, V_n$ est inversible et la matrice $P^{-1}AP$ est la matrice diagonale $D = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ dont les coefficients sont $\lambda_1, \lambda_2, \dots, \lambda_n$.

  Posons $X = PY$. Les solutions du système différentiel $Y' = DY$ sont

  $$Y(t) = \begin{bmatrix} k_1 e^{\lambda_1 t} \\ k_2 e^{\lambda_2 t} \\ \vdots \\ k_n e^{\lambda_n t} \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

  où $K = \begin{bmatrix} k_1 \\ \vdots \\ k_n \end{bmatrix}$ est un vecteur constant quelconque. Pour trouver les solutions du système $X' = AX$, on doit effectuer le produit matriciel $X = PY$; calculons

  $$R(t) = \Big[V_1 \ V_2 \ \cdots \ V_n \Big] \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} = \Big[e^{\lambda_1 t} V_1 \ e^{\lambda_2 t} V_2 \ \cdots \ e^{\lambda_n t} V_n \Big]$$

  Puisque $P = [V_1 \ \cdots \ V_n]$, en multipliant à droite par le vecteur K, on obtient

  $$X(t) = PY(t) = R(t)K = k_1 e^{\lambda_1 t} V_1 + k_2 e^{\lambda_2 t} V_2 + \cdots + k_n e^{\lambda_n t} V_n$$

  où $k_1, \dots, k_n$ sont des nombres quelconques.

- Les solutions $e^{\lambda_1 t} V_1, \dots, e^{\lambda_n t} V_n$ s'appellent des *solutions propres* : ce sont les colonnes de la matrice $R(t)$.
- La solution générale est une combinaison linéaire des solutions propres. Puisque les vecteurs propres V_i sont indépendants, les solutions propres sont indépendantes.

  Les solutions propres forment une base de l'espace vectoriel des solutions du système. L'espace des solutions est donc de dimension n.

- La matrice $R(t)$ est inversible quel que soit t.

A est trigonalisable. On peut trouver une matrice inversible P telle que $T = P^{-1}AP$ est triangulaire, ou même triangulaire par blocs ; les coefficients diagonaux de T sont les valeurs propres de A. On résout le système différentiel $Y' = TY$ en prenant les équations successivement, comme dans l'exemple 2 page 482 ; on trouve n solutions $Y_1, Y_2 \dots, Y_n$ indépendantes dont les coordonnées sont de la forme $q(t)e^{\lambda t}$, où $q(t)$ est une fonction polynôme et λ une valeur propre de A. On dispose $Y_1(t), \dots, Y_n(t)$ en colonnes, on calcule la matrice carrée $R(t) = P\Big[Y_1(t) \ Y_2(t) \ \cdots \ Y_n(t) \Big]$

et les solutions sont les fonctions

$$X(t) = R(t)K \text{ , où } K \text{ est un vecteur de constantes quelconques.}$$

Les colonnes de $R(t)$ forment une base de l'espace vectoriel des solutions.

Les valeurs propres de A ne sont pas réelles. Soit λ une valeur propre complexe de A et V un vecteur propre associé. Posons

$$\lambda = a + \boldsymbol{i}\,\omega \quad \text{et} \quad V = U_1 + \boldsymbol{i}\,U_2$$

où a, ω sont des nombres réels, $\omega \neq 0$ et U_1, U_2 des vecteurs à coefficients réels. On a

$$AV = \lambda V = (a + \boldsymbol{i}\,\omega)(U_1 + \boldsymbol{i}\,U_2) = (aU_1 - \omega U_2) + \boldsymbol{i}\,(\omega U_1 + aU_2)$$

donc en séparant partie réelle et partie imaginaire, il vient

$$AU_1 = aU_1 - \omega U_2 \qquad , \qquad AU_2 = \omega U_1 + aU_2$$

Supposons A de taille 2.

Les vecteurs U_1, U_2 étant indépendants (car $\omega \neq 0$), la matrice de la transformation $Z \mapsto AZ$ dans la base (U_1, U_2) de $\mathbb{R}^2$ est $C = \begin{bmatrix} a & \omega \\ -\omega & a \end{bmatrix}$. Par conséquent, on a l'égalité $C = P^{-1}AP$, où $P = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$.

En posant $X = PY$, on sait que les fonctions $Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}$ sont solutions de $Y' = CY$, c'est-à-dire

$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = Y' = CY = \begin{bmatrix} a & \omega \\ -\omega & a \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} ay_1 + \omega y_2 \\ -\omega y_1 + ay_2 \end{bmatrix}$$

Introduisons la fonction $z = y_1 + \boldsymbol{i}\,y_2$. L'égalité vectorielle ci-dessus s'écrit

$$z' = y_1' + \boldsymbol{i}\,y_2' = (ay_1 + \omega y_2) + \boldsymbol{i}\,(-\omega y_1 + ay_2) = (a - \boldsymbol{i}\,\omega)(y_1 + \boldsymbol{i}\,y_2) = \overline{\lambda} z$$

d'où $z(t) = ke^{\overline{\lambda} t}$, avec $k = k_1 + \boldsymbol{i}\,k_2$ un nombre complexe quelconque (remarque page 444). Puisque $e^{\overline{\lambda} t} = e^{at}e^{-\boldsymbol{i}\,\omega t}$, On obtient ainsi

$$z(t) = (k_1 + \boldsymbol{i}k_2)e^{at}\big(\cos\omega t - \boldsymbol{i}\sin\omega t\big) = e^{at}\big[(k_1\cos\omega t + k_2\sin\omega t) + \boldsymbol{i}(-k_1\sin\omega t + k_2\cos\omega t)\big]$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = e^{at} \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

et en posant $K = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$, il vient finalement

$$(*) \qquad X(t) = e^{at}\begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix} K = R(t)K$$

où K est un vecteur constant quelconque. Les colonnes de $R(t)$ forment une base de l'espace vectoriel des solutions : cet espace est donc de dimension 2.

Supposons maintenant que A est de de taille n quelconque et diagonalisable sur $\mathbb{C}$. Procédons comme ci-dessus pour chaque vecteur propre complexe $U_1 + \boldsymbol{i}\,U_2$, $V_1 + \boldsymbol{i}\,V_2$, etc écrivons les vecteurs $U_1, U_2, V_1, V_2, \ldots$ en colonne et complétons par les vecteurs propres réels : on forme ainsi une matrice carrée P inversible. Par le changement d'inconnue

$X = PY$, le système devient $Y' = CY$, avec

$$C = \begin{bmatrix} a & \omega & 0 & 0 & 0 & \cdots & 0 \\ -\omega & a & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & c & \varphi & 0 & \cdots & 0 \\ 0 & 0 & -\varphi & c & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & \\ 0 & \cdots & & & r_1 & & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & & & 0 & & r_m \end{bmatrix}$$

$a + \boldsymbol{i}\omega$, $c + \boldsymbol{i}\varphi$, etc étant les valeurs propres non réelles et $r_1, \ldots, r_m$ les valeurs propres réelles de A. On résout le système $Y' = CY$ bloc par bloc et les solutions sont les fonctions $X(t) = PY(t) = R(t)K$, où K est un vecteur formé de n constantes réelles arbitraires. L'espace des solutions est encore de dimension n.

Exemple 3. Considérons le système différentiel (S) $\begin{cases} x' = 2y - 6 \\ y' = x - y + 2 \end{cases}$

Une solution constante est déterminée par $x' = 2y - 6 = 0$ et $y' = x - y + 2 = 0$, donc $y = 3, x = 1$ et la fonction constante $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ est solution. Puisque la différence de deux solutions est une solution du système homogène, posons

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} u+1 \\ v+3 \end{bmatrix}$$

On a $u' = x' = 2y - 6 = 2v$ et $v' = y' = x - y + 2 = (u+1) - (v+3) + 2 = u - v$, d'où le système différentiel en u, v :

$$(h) \qquad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Appelons A la matrice du système (h).

Valeurs propres de A : ce sont les racines du polynôme caractéristique

$$\begin{vmatrix} -z & 2 \\ 1 & -1-z \end{vmatrix} = -z(-1-z) - 2 = z^2 + z - 2 = (z-1)(z+2)$$

Il y a deux valeurs propres distinctes 1 et -2, donc la matrice A est diagonalisable.

Vecteurs propres de A : Un vecteur propre $V_1 = \begin{bmatrix} a \\ b \end{bmatrix}$ pour la valeur propre 1 est solution de $(A - I_2)V_1 = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 0$, donc $V_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ est vecteur propre pour la valeur propre 1. Pour la valeur propre -2, on a $A + 2I_2 = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$, donc $V_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ est vecteur propre.

Les solutions de (h) sont

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix} = k_1 e^t \begin{bmatrix} 2 \\ 1 \end{bmatrix} + k_2 e^{-2t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \text{ avec } k_1 \text{ et } k_2 \text{ quelconques,}$$

et les solutions de (S) sont $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = k_1 e^t \begin{bmatrix} 2 \\ 1 \end{bmatrix} + k_2 e^{-2t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, c'est-à-dire

$$\begin{cases} x(t) &= 2k_1 e^t + k_2 e^{-2t} + 1 \\ y(t) &= k_1 e^t - k_2 e^{-2t} + 3 \end{cases}$$

Il y a une seule solution de conditions initiales $x(t_0)=x_0, y(t_0)=y_0$: on la trouve en calculant les constantes k_1 et k_2 telles que $x_0=2k_1 e^{t_0}+k_2 e^{-2t_0}+1$ et $y_0=k_1 e^{t_0}-k_2 e^{-2t_0}+3$.

Dessin des trajectoires. Quand t parcourt $\mathbb{R}$, le point $M(t) = \big(x(t), y(t)\big)$ décrit une courbe passant par le point initial $M_0 = (x_0, y_0)$: c'est la *trajectoire* de M_0.
Si la condition initiale est (1,3), la solution est la fonction constante $\big(x(t),y(t)\big)=(1,3)$. La trajectoire du point $E = (1, 3)$ est donc réduite au point E : Le point E est un équilibre du système.
Pour dessiner les différentes trajectoires, plaçons-nous dans le repère d'origine E et d'axes dirigés par les vecteurs propres V_1, V_2. On a $\overrightarrow{EM(t)} = k_1 e^t V_1 + k_2 e^{-2t} V_2$ et les coordonnées de $M(t)$ dans ce repère sont $\alpha = k_1 e^t, \beta = k_2 e^{-2t}$.

- Si $k_1 = 0$ et $k_2 \neq 0$, alors $M(t)$ est sur l'axe dirigé par V_2, β garde le signe de k_2 et tend vers 0 exponentiellement. Si $k_2 > 0$, la trajectoire est la demi-droite d'origine E dirigée par V_2 et parcourue vers E ; si $k_2 < 0$, c'est la demi-droite opposée, parcourue également vers E (figure 1).
- Si $k_2 = 0$ et $k_1 \neq 0$, alors $\beta = 0$ et α garde le signe de k_1. Si $k_1 > 0$, la trajectoire est la demi-droite d'origine E dirigée par V_1 et parcourue en s'éloignant vers l'infini ; si $k_1 < 0$, c'est la demi-droite opposée (figure 1).
- Supposons $k_1 k_2 \neq 0$. Pour trouver l'équation des trajectoires, éliminons t entre les égalités $\alpha = k_1 e^t, \beta = k_2 e^{-2t}$. Il vient $\alpha^2 \beta = k_1^2 e^{2t} k_2 e^{-2t} = k_1^2 k_2$. La courbe d'équation $\beta = c/\alpha^2$, où c est une constante non nulle, est formée de deux branches ayant les axes pour asymptotes (figure 2). Chaque branche est une trajectoire (quand t varie, les coordonnées α et β gardent un signe constant).

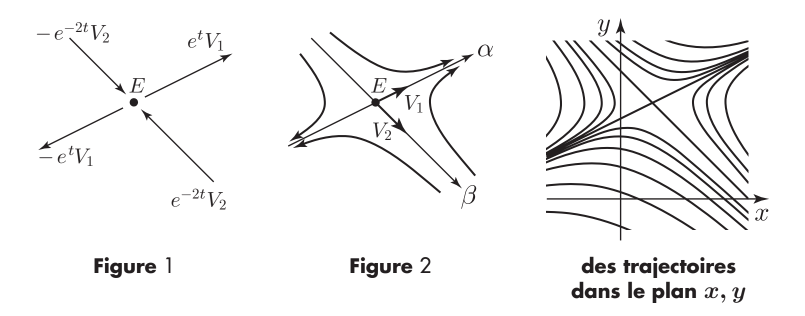


Figure 1 — **Figure** 2 — **des trajectoires dans le plan x, y**

Exemple 4 : un modèle de compétition végétale. Deux espèces végétales (e_1) et (e_2) sont en compétition dans un périmètre donné. Le nombre d'individus de chaque espèce fluctue autour d'une valeur moyenne, avec des écarts $x(t)$ et $y(t)$

satisfaisant le système différentiel

$$\begin{cases} \dot{x} = -(1/4)(x-y) \\ \dot{y} = -(1/4)(25x+7y) \end{cases} \quad \text{de matrice } A = \begin{bmatrix} -1/4 & 1/4 \\ -25/4 & -7/4 \end{bmatrix}.$$

À la suite d'une perturbation de l'écosystème, un recensement à l'instant $t=0$ montre que $x(0) = x_0$ et $y(0) = 0$. Comment vont évoluer $x(t)$ et $y(t)$?

Le polynôme caractéristique de A est z^2+2z+2, les valeurs propres sont $-1+\boldsymbol{i}$, $-1-\boldsymbol{i}$ et $V=\begin{bmatrix} 1 \\ -3+4\,\boldsymbol{i} \end{bmatrix}=\begin{bmatrix} 1 \\ -3 \end{bmatrix}+\boldsymbol{i}\begin{bmatrix} 0 \\ 4 \end{bmatrix}$ est vecteur propre pour $-1+\boldsymbol{i}$. D'après la formule $(*)$ page 484, les solutions sont les fonctions $X(t)=R(t)K$, où K est un vecteur constant et

$$R(t) = e^{-t} \begin{bmatrix} 1 & 0 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

On a $R(0) = \begin{bmatrix} 1 & 0 \\ -3 & 4 \end{bmatrix}$ donc $X(0) = \begin{bmatrix} 1 & 0 \\ -3 & 4 \end{bmatrix} K$. Pour que $x(0) = x_0$ et $y_0 = 0$, il faut prendre le vecteur $K = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$ tel que $\begin{cases} k_1 = x_0 \\ -3k_1 + 4k_2 = 0 \end{cases}$, c'est-à-dire $k_1 = x_0$ et $k_2 = (3/4)x_0$. On obtient ainsi

$$X(t) = e^{-t} \begin{bmatrix} 1 & 0 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} x_0 \\ (3/4)x_0 \end{bmatrix} = \frac{x_0}{4} e^{-t} \begin{bmatrix} 1 & 0 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 4\cos t+3\sin t \\ -4\sin t+3\cos t \end{bmatrix}$$

c'est-à-dire $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \frac{x_0}{4} e^{-t} \begin{bmatrix} 4\cos t + 3\sin t \\ -25\sin t \end{bmatrix}$. Les fonctions $x(t)$ et $y(t)$ tendent vers 0 quand t tend vers l'infini : cela veut dire que les populations végétales reviennent à leur proportion d'équilibre. La figure 1 montre la trajectoire dans le plan des (x, y) ; sur la figure 2, on a représenté les graphes des fonctions $x(t)$ et $y(t)$ (t est en abscisse).

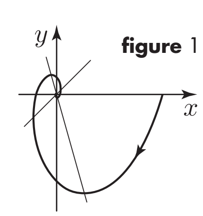


figure 1

Dans la première équation du système différentiel, on voit que $\dot{x}$ s'annule quand $x-y=0$: les valeurs extrêmes de $x(t)$ sont donc obtenues quand les deux populations ont le même nombre d'individus. Cela correspond, sur la trajectoire, aux points à tangente verticale : ils sont sur la droite d'équation $y = x$. De même, les points de la trajectoire où la tangente est horizontale sont caractérisés par $\dot{y} = 0$: d'après la seconde équation du système, ils sont sur la droite d'équation $y = -(25/7)x$.

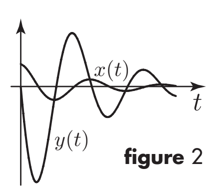


figure 2

1.2 Matrice résolvante

Nous venons de voir que la solution générale d'un système différentiel $X' = AX$ de taille n est toujours combinaison linéaire de n solutions indépendantes $X_1(t), X_2(t), \ldots, X_n(t)$. Si l'on dispose ces solutions en colonnes, on obtient la matrice $R(t) = \big[X_1(t) \cdots X_n(t)\big]$ et la solution générale du système est $X(t) = R(t)K$, où K est un vecteur constant quelconque.

Une telle matrice $R(t)$ s'appelle une *matrice résolvante* du système.

► La matrice $R(t)$ est inversible quel que soit t.
► On a $R'(t) = AR(t)$ quel que soit t.

En effet $R'(t) = \left[X_1'(t)\ X_2'(t)\ \ldots\ X_n'(t)\right] = \left[AX_1(t)\ AX_2(t)\ \ldots\ AX_n(t)\right] = AR(t)$.

Solution avec condition initiale

En pratique, on cherche souvent la solution $X(t)$ prenant en $t = t_0$ une valeur vectorielle $X(t_0) = X_0$ donnée. Pour que $X(t_0) = X_0$, il faut et il suffit que l'on ait $R(t_0)K = X_0$, ou encore $K = R(t_0)^{-1}X_0$; on a donc $X(t) = R(t)R(t_0)^{-1}X_0$.

La solution telle que $X(t_0) = X_0$ *est* $X(t) = R_{t_0}(t)X_0$, *où* $R_{t_0}(t) = R(t)R(t_0)^{-1}$.

On a $R_{t_0}(t_0) = I_n$. Prenons $t_0 = 0$ et soit $X(t) = R_0(t)X_0$ la solution telle que $X(0) = X_0$. Si a est un nombre réel, la fonction $Y(t) = X(t + a)$ est solution, car $Y'(t) = X'(t+a) = AX(t+a) = AY(t)$. Puisque $Y(0) = X(a)$, la fonction $Y(t)$ est la solution telle que $Y(0) = X(a)$, donc $Y(t) = R_0(t)X(a)$. Puisque $X(t+a) = R_0(t+a)X_0$, il vient $R_0(t + a)X_0 = R_0(t)X(a) = R_0(t)R_0(a)X_0$ et comme cela est vrai quel que soit le vecteur initial X_0, on en déduit l'identité matricielle

$$R_0(t + a) = R_0(t)R_0(a) \text{ , quels que soient } t \text{ et } a.$$

En particulier $I_n = R_0(t - t) = R_0(t)R_0(-t)$, donc $R_0(t)^{-1} = R_0(-t)$. Il s'ensuit que pour tout entier k positif ou négatif, on a $R_0(kt) = \left[R_0(t)\right]^k$.
La fonction $t \mapsto R_0(t)$ transforme somme en produit et opposé en inverse, comme une exponentielle.

Recherche de la matrice du système à partir des solutions

Supposons que le système différentiel $X' = AX$ est de taille n. Si l'on en connaît n solutions indépendantes $X_1, X_2, \ldots, X_n$, la matrice $R(t)$ dont les colonnes sont $X_1(t), X_2(t), \ldots X_n(t)$ est une résolvante. Puisque $R'(t) = AR(t)$, on en déduit $A = R'(t)R(t)^{-1}$.

Application : modélisation d'une diffusion compartimentale

Une substance présente dans deux compartiments tissulaires (1) et (2) peut diffuser de l'un vers l'autre à un taux proportionnel à la quantité présente dans le compartiment source (loi d'action de masse). Soient $x_1(t)$ et $x_2(t)$ les quantités de substance présentes à l'instant t dans les compartiments (1) et (2). Ces fonctions sont solutions d'un système différentiel homogène (comparer à l'exercice 11 page 197)

$$(h) \qquad \begin{cases} \dot{x} = ax + by \\ \dot{y} = cx + dy \end{cases}$$

Dans la pratique, les coefficients de diffusion a,b,c,d sont inconnus. En effectuant dans les compartiments des dosages à des instants $t_1, t_2, \ldots$, on peut faire une estimation numérique des fonctions x et y. En général, x et y décroissent exponentiellement,

de sorte qu'on propose pour ces fonctions des formules du type

$$\begin{cases} x(t) &= p_1 e^{\lambda_1 t} + p_2 e^{\lambda_2 t} \\ y(t) &= q_1 e^{\lambda_1 t} + q_2 e^{\lambda_2 t} \end{cases}, \quad \text{où } \lambda_1, \lambda_2, p_1, p_2, q_1, q_2 \text{ sont des constantes.}$$

Il s'agit bien d'une forme possible pour une solution d'un système différentiel linéaire homogène.

Procédons comme dans l'exemple page 305. En faisant varier les quantités initiales, on calcule des valeurs approchées pour les exposants λ_1 et λ_2 et pour les coefficients p_1, p_2, q_1, q_2 : on peut ainsi obtenir deux fonctions de diffusion $X_1(t) = \begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix}$ et $X_2(t) = \begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix}$ non proportionnelles. Formons la matrice $R(t) = \begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix}$.

Les fonctions X_1 et X_2 sont solutions du système différentiel linéaire $\dot{X} = AX$ de matrice $A = \dot{R}(t)R(t)^{-1}$ ($\dot{R}$ est la matrice dérivée par rapport à t). Les coefficients de A doivent être indépendants de t, mais on a intérêt à faire le calcul en prenant différentes valeurs de t et à retenir un résultat moyen.

- Les coefficients de la matrice A sont des estimations numériques des coefficients de diffusion a, b, c, d dans (h).
- Une fois calculée la matrice A, on dispose du système différentiel $\dot{X} = AX$ qui permet de résoudre le problème des diffusions pour n'importe quelle condition initiale.
- On connaît les valeurs propres de A : ce sont les coefficients de t dans les exponentielles qui composent $X_1(t)$ ou $X_2(t)$.

Par exemple, supposons qu'on a fait les deux estimations

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = \begin{bmatrix} 0{,}148\, e^{-0{,}18\, t} + 4{,}85\, e^{-0{,}5\, t} \\ 0{,}652\, e^{-0{,}18\, t} + 2{,}34\, e^{-0{,}5\, t} \end{bmatrix} \quad \text{et} \quad \begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 0{,}528\, e^{-0{,}18\, t} + 3{,}47\, e^{-0{,}5\, t} \\ 2{,}32\, e^{-0{,}18\, t} + 1{,}68\, e^{-0{,}5\, t} \end{bmatrix}$$

Formons la matrice $R(t) = \begin{bmatrix} x_1(t) & x_2(t) \\ y_1(t) & y_2(t) \end{bmatrix}$ et calculons sa dérivée en $t = 0$ par exemple :

$$R(t) = \begin{bmatrix} 0{,}148\, e^{-0{,}18\, t} + 4{,}85\, e^{-0{,}5\, t} & 0{,}528\, e^{-0{,}18\, t} + 3{,}47\, e^{-0{,}5\, t} \\ 0{,}652\, e^{-0{,}18\, t} + 2{,}34\, e^{-0{,}5\, t} & 0{,}528\, e^{-0{,}18\, t} + 3{,}47\, e^{-0{,}5\, t} \end{bmatrix}, \quad \dot{R}(0) = \begin{bmatrix} -2{,}45 & -1{,}83 \\ -1{,}28 & -1{,}25 \end{bmatrix}$$

La matrice du système est $A = \dot{R}(0)R(0)^{-1} = \begin{bmatrix} -0{,}54 & -0{,}08 \\ -1{,}72 & -1{,}42 \end{bmatrix}$. Les valeurs propres de A sont $-0{,}18$ et $-0{,}5$.

1.3 Principaux types de trajectoires

La forme des trajectoires d'un système différentiel $X' = AX$ dépend des valeurs propres de A.

- L'origine est un point d'équilibre : il constitue une trajectoire.
- Si λ est une valeur propre réelle avec vecteur propre V, la trajectoire associée à la solution propre $e^{\lambda t}V$ est la demi-droite passant par l'origine et dirigée par V. La demi-droite opposée est la trajectoire associée à la solution $-e^{\lambda t}V$.

Stabilité de l'équilibre

L'origine O est un *équilibre stable* si toutes les solutions de condition initiale $X(t_0)$ voisines de zéro restent proches de O pour $t \geqslant t_0$.
D'après les différents types de solutions trouvés lors de la résolution, on a le résultat suivant.

Conditions de stabilité

- *Si toutes les valeurs propres de la matrice A ont leur partie réelle strictement négative, alors toutes les solutions de $X' = AX$ tendent vers 0 quand t tend vers $+\infty$ et l'origine est un équilibre stable.*
- *Si l'une au moins des valeurs propres est de partie réelle strictement positive, l'origine est un équilibre instable.*
- *Si la matrice A est diagonalisable et si toutes ses valeurs propres sont de partie réelle négative ou nulle, alors l'origine est un équilibre stable.*

Démonstration. Chaque valeur propre réelle λ détermine une solution propre $U(t) = e^{\lambda t}V$, où V un vecteur propre associé, de norme $\|U(t)\| = e^{\lambda t}\|V\|$. Si $\lambda = 0$, la norme de $U(t)$ est constante. Si $\lambda < 0$, la norme tend vers 0 en décroissant quand t tend vers $+\infty$. Si $\lambda > 0$, la norme tend vers $+\infty$ quand t tend vers $+\infty$ et l'équilibre n'est pas stable.
Si $\lambda = a + \boldsymbol{i}\,\omega$ est une valeur propre complexe, les solutions associées sont de la forme $U(t) = e^{at}\begin{bmatrix} c_1 \cos(\omega t + \varphi_1) \\ c_2 \cos(\omega t + \varphi_2) \end{bmatrix}$ et l'on a $\|U(t)\| \leqslant ke^{at}$, où k est une constante ; on en déduit que si $a \leqslant 0$, la norme de $U(t)$ reste bornée par k quand t parcourt $[0, +\infty[$ et que si $a < 0$, la norme tend vers 0 quand t tend vers $+\infty$.
Si A est diagonalisable, toute solution est combinaison linéaire de solutions du type précédent, d'où les affirmations ci-dessus dans ce cas.
Quand A est seulement trigonalisable, les solutions peuvent avoir dans leur coordonnées des fonctions de la forme $q(t)e^{at}$ ou $q(t)e^{at}\cos(\omega t + \varphi)$, où a est une partie réelle de valeur propre et $q(t)$ un polynôme. Si $a < 0$, alors $\lim\limits_{t \to +\infty} q(t)e^{at} = 0$, donc ces fonctions restent bornées quand t parcourt $[0, +\infty[$. On en déduit que si toutes les valeurs propres sont de partie réelle strictement négative, alors toutes les solutions tendent vers O et l'équilibre est stable. ■

Critère de stabilité en dimension 2. *Supposons que la matrice A est de taille 2 et possède au moins une valeur propre non nulle. Alors pour le système différentiel $X' = AX$, l'équilibre O est stable si et seulement si les coefficients du polynôme caractéristique de A sont positifs ou nuls.*

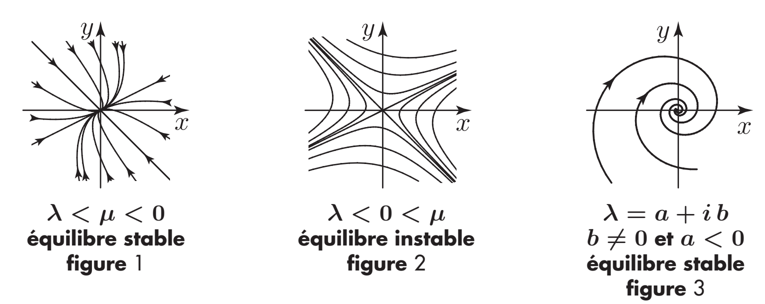
Démonstration. Pour que l'équilibre $(0, 0)$ soit stable, il faut que les valeurs propres soient de partie réelle négative ou nulle. Réciproquement, si les valeurs propres λ et μ sont de partie réelle strictement négative, ou si $\lambda < \mu = 0$, l'équilibre est stable ; si leur partie réelle est nulle, alors λ et μ sont imaginaires pures (on a exclu le cas $\lambda = \mu = 0$) et l'équilibre est stable, car les coordonnées des solutions sont des fonctions $a\cos(\omega t + \varphi)$. Ainsi l'équilibre est stable si et seulement si λ et μ sont de partie réelle négative ou nulle. Le polynôme caractéristique de A est $P = (z - \lambda)(z - \mu) = z^2 + pz + q$, où $p = -(\lambda + \mu)$ et $q = \lambda\mu$. Quand les racines sont réelles, elles sont toutes deux négatives ou nulles à la condition que leur produit q soit positif ou nul

et que leur somme $-p$ soit négative ou nulle ; quand les racines sont complexes conjuguées, on a $q = \lambda\bar{\lambda} > 0$ et la condition de stabilité est $-p = 2\,\mathscr{Re}(\lambda) \leqslant 0$. ■

Supposons encore la matrice A de taille 2.

- Si les valeurs propres sont réelles non nulles et de même signe, les trajectoires non rectilignes sont toutes tangentes à l'une des directions propres (figure 1).
- Si les valeurs propres sont réelles et de signes contraires, les trajectoires non rectilignes sont des « branches hyperboliques » asymptotes aux directions propres (fig. 2).
- Si les valeurs propres sont non réelles (donc complexes conjuguées), les trajectoires sont des spirales (figure 3), ou des ellipses dans le cas de valeurs propres imaginaires pures (voir figure page 499).

Voici différentes possibilités de trajectoires, selon les valeurs propres λ et μ :



$\boldsymbol{\lambda < \mu < 0}$
équilibre stable
figure 1

$\boldsymbol{\lambda < 0 < \mu}$
équilibre instable
figure 2

$\boldsymbol{\lambda = a + i\,b}$
$\boldsymbol{b \neq 0}$ **et** $\boldsymbol{a < 0}$
équilibre stable
figure 3

Si l'on change le signe de toutes les valeurs propres, les trajectoires sont les mêmes mais leur sens de parcours est inversé.

On retrouve les dessins obtenus pages 187 et 188 dans l'étude des itérations linéaires. Ce n'est pas un hasard. En effet, si l'on choisit un pas $h>0$, les itérés de X_0 par la transformation $X \mapsto R_0(h)X$ sont les points $X_n = [R_0(h)]^n X_0$. Puisqu'on a $[R_0(h)]^n = R_0(nh)$ d'après les propriétés de la résolvante R_0, il vient $X_n = R_0(nh)X_0$, donc les points X_n sont sur la trajectoire de X_0. Si par exemple A est la matrice diagonale de valeurs propres λ et μ, alors $R_0(h)$ est diagonale de valeurs propres $e^{\lambda h}, e^{\mu h}$. Dans le cas $\lambda<\mu<0$, on a $0<e^{\lambda h}<e^{\mu h}<1$ et la trajectoire a bien la forme des itérés d'un point par une matrice de valeurs propres strictement comprises entre 0 et 1. On retrouve de même les autres formes de trajectoires, selon la position de $e^{\lambda h}$ et de $e^{\mu h}$ par rapport à 1, c'est-à-dire selon le signe de λ et μ.

1.4 Système avec second membre

Pour résoudre un système différentiel linéaire $(s)\quad X' = AX + B(t)$ où le second membre $B(t)$ est un vecteur de fonctions,

a) on résout le système homogène $(h)\;\; X' = AX$: les solutions de (h) s'écrivent $X(t) = R(t)K$, où $R(t)$ est une matrice résolvante et K un vecteur constant quelconque ;

b) et l'on cherche une solution particulière $S(t)$ du système complet.

Les solutions de (s) sont toutes les fonctions $R(t)K + S(t)$, où K est un vecteur constant quelconque (propriété page 481).

Dans les applications, le problème possède parfois une solution particulière évidente, par exemple une solution constante. Mais nous allons voir que l'on peut calculer une solution particulière de (s) au moyen des solutions de (h).

Méthode de variation de la constante

Cherchons une solution $S(t)$ de la forme $S(t) = R(t)K(t)$, où $K(t)$ est une fonction inconnue : c'est la *méthode de variation de la constante*, analogue à celle pratiquée page 444 pour une équation linéaire numérique.

On a $S' = R'K + RK'$, car chaque coefficient de $S(t)$ est une somme de termes $r(t)k(t)$, où $r(t)$ et $k(t)$ sont des coefficients de R et K. La condition pour que $S(t)$ soit solution est donc

$$\begin{aligned} S' = AS + B(t) &\iff R'K + RK' = ARK + B \\ &\iff ARK + RK' = ARK + B \text{ , car } R' = AR \end{aligned}$$

et en simplifiant par ARK, il vient $RK' = B$, ou encore $K' = R^{-1}B$.

$S(t) = R(t)K(t)$ *est solution de* (s) *si et seulement si* $K'(t) = R(t)^{-1}B(t)$.

On choisit *une* primitive de la fonction vectorielle $R(t)^{-1}B(t)$ en intégrant coefficient par coefficient et la solution générale de (s) s'écrit

$$X(t) = R(t)K + R(t)\left[\int^t R(s)^{-1}B(s)\,ds\right], \text{ avec } K \text{ un vecteur de constantes.}$$

Exemple. Reprenons l'exemple de la diffusion d'une substance dans deux compartiments tissulaires (page 488) en supposant que la matrice du système est

$$A = \begin{bmatrix} -2 & 1/4 \\ 3 & -1 \end{bmatrix}$$

Injectons la substance dans le compartiment (1) avec un débit v. Les quantités $x(t)$ et $y(t)$ présentes dans les compartiments (1) et (2) sont alors solutions du système

$$(s) \qquad \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} vt \\ 0 \end{bmatrix}$$

Le polynôme caractéristique de A est $\begin{vmatrix} -2-z & 1/4 \\ 3 & -1-z \end{vmatrix} = z^2+3z+(5/4) = (z+5/2)(z+1/2)$, les valeurs propres sont $-5/2$ et $-1/2$ et les vecteurs propres sont $V_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ et $V_2 = \begin{bmatrix} 1 \\ 6 \end{bmatrix}$. La matrice résolvante du système homogène $X' = AX$ est donc

$$R(t) = \begin{bmatrix} V_1 e^{-(5/2)t} & V_2 e^{-(1/2)t} \end{bmatrix} = \begin{bmatrix} e^{-(5/2)t} & e^{-(1/2)t} \\ -2e^{-(5/2)t} & 6e^{-(1/2)t} \end{bmatrix}$$

On a $R(t)^{-1} = \dfrac{1}{8} \begin{bmatrix} 6e^{(5/2)t} & -e^{(5/2)t} \\ 2e^{(1/2)t} & e^{(1/2)t} \end{bmatrix}$. Le second membre est $B(t) = \begin{bmatrix} vt \\ 0 \end{bmatrix}$ et

$$K'(t) = R(t)^{-1}B(t) = \frac{v}{4} \begin{bmatrix} 3te^{(5/2)t} \\ te^{(1/2)t} \end{bmatrix}$$

On a $\int te^{(5/2)t}\,dt = \dfrac{2}{25}(5t-2)e^{(5/2)t}$ et $\int te^{(1/2)t}\,dt = 2(t-2)e^{(1/2)t}$, d'où $K(t) = \begin{bmatrix} (3v/50)(5t-2)e^{(5/2)t} \\ (v/2)(t-2)e^{(1/2)t} \end{bmatrix}$. Une solution particulière de (s) est donc

$$S(t) = R(t)K(t) = \frac{4v}{25}\begin{bmatrix} 5t-7 \\ 15t-36 \end{bmatrix}$$

La solution générale de (s) est

$$X(t) = R(t)\begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + S(t) = \begin{bmatrix} k_1e^{-(5/2)t} + k_2e^{-(1/2)t} + \dfrac{4v(5t-7)}{25} \\ -2k_1e^{-(5/2)t} + 6k_2e^{-(1/2)t} + \dfrac{4v(15t-36)}{25} \end{bmatrix}$$

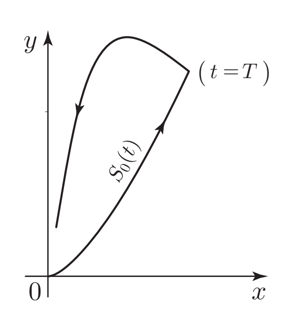
avec k_1 et k_2 des constantes arbitraires.

Supposons qu'à l'instant initial $t=0$ où l'on commence l'injection, les compartiments ne contiennent pas de substance. Les conditions initiales sont donc $x(0) = y(0) = 0$. En calculant les constantes k_1 et k_2 pour satisfaire ces conditions, on obtient

$$S_0(t) = \begin{bmatrix} x_0(t) \\ y_0(t) \end{bmatrix} = v\begin{bmatrix} (3/25)e^{-(5/2)t}+e^{-(1/2)t}+(4/5)t-28/25 \\ -(6/25)e^{-(5/2)t}+6e^{-(1/2)t}+(12/5)t-144/25 \end{bmatrix}$$

Supposons qu'on arrête l'injection à l'instant T : passé cet instant, les fonctions $x(t)$ et $y(t)$ sont régies par le système homogène $X' = AX$. Pour $t > T$, $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$ est donc la solution $S_1(t)$ du système homogène telle que $S_1(T) = S_0(T)$. On a $S_1(t) = R(t)C$, où le vecteur constant C doit vérifier $R(T)C = S_0(T)$; en résolvant cette équation linéaire, on obtient C, ce qui permet de calculer $S_1(t)$.

La figure ci-contre montre l'allure de la courbe $\big(x(t), y(t)\big)$ ainsi obtenue. La partie inférieure de la courbe correspond à $S_0(t)$: tant que $t < T$, les quantités $x(t)$ et $y(t)$ augmentent. À l'instant T, il y a rebroussement : pour $t > T$, $x(t)$ décroît tandis que $y(t)$ continue à croître jusqu'à un maximum, puis $x(t)$ et $y(t)$ décroissent vers 0 (partie supérieure de la courbe).



Comme $y(t)$ continue à croître un moment après l'arrêt de l'injection, il faut interrompre celle-ci à temps pour ne pas dépasser la concentration maximale admissible dans le compartiment.

Application à une équation linéaire du second ordre

Considérons l'équation différentielle linéaire (e) $x'' + px' + qx = b(t)$, où p et q sont des constantes. En posant $y = x'$, il vient $y' = x'' = -px' - qx + b(t)$ et le système différentiel suivant est équivalent à (e) :

$$(s) \qquad \begin{cases} x' = y \\ y' = -qx - py + b(t) \end{cases}$$

Si l'on pose $X = \begin{bmatrix} x \\ y \end{bmatrix}$, ce système s'écrit sous forme matricielle

$$X' = AX + B(t), \text{ avec } A = \begin{bmatrix} 0 & 1 \\ -q & -p \end{bmatrix} \text{ et } B(t) = \begin{bmatrix} 0 \\ b(t) \end{bmatrix}.$$

La résolution de (s) fournit les solutions $x(t)$ de (e). En particulier, la méthode de variation de la constante présentée ci-dessus permet toujours d'exprimer une solution particulière de (e) : les calculs conduisent à la formule page 451, où w est ici de la forme $w_0 e^{-pt}$.

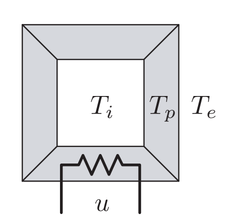
2. Système différentiel linéaire contrôlé

Considérons une quantité vectorielle $X(t)$ régie par un système différentiel linéaire homogène $X' = AX$ de taille n. Supposons qu'on puisse modifier le système au moyen de paramètres de commande $u_1, u_2, \ldots$ agissant en continu.

Exemple. Un four électrique est constitué d'une enceinte entourée d'une paroi chauffante épaisse. Notons

- c_i et c_p les capacités calorifiques de l'intérieur et de la paroi du four,
- s_i et s_e les aires des parois intérieures et extérieures,
- r_i et r_e les coefficients de radiation de la paroi intérieure et de la paroi extérieure.

Les variations de température par convection sont proportionnels à la surface d'échange et à l'écart des températures. Notons T_p la température moyenne dans la paroi, T_e et T_i les températures à l'extérieur et à l'intérieur du four. Si u est le flux de chaleur apporté par le chauffage du four, les bilans caloriques conduisent aux équations suivantes :



$$\begin{aligned} c_p \dot{T}_p &= -s_e r_e (T_p - T_e) - s_i r_i (T_p - T_i) + u \qquad \text{(pour la paroi)} \\ c_i \dot{T}_i &= s_i r_i (T_p - T_i) \qquad \text{(pour l'intérieur).} \end{aligned}$$

Introduisons les excès de température $x = T_p - T_e$ et $y = T_i - T_e$ par rapport à l'extérieur. Puisque $T_p - T_i = x - y$, il vient

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -(s_e r_e + s_i r_i)/c_p & s_i r_i / c_p \\ s_i r_i / c_i & -s_i r_i / c_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1/c_p \\ 0 \end{bmatrix} u$$

On peut régler u au moyen d'un dispositif électrique : u est le *paramètre de contrôle* et l'on dit que le système différentiel est *contrôlé*. La question est la suivante :

> par un réglage en continu $u(t)$ convenable, peut-on, à partir de n'importe quelle température initiale, atteindre dans le four une température donnée quelconque ?

Plus généralement, considérons un système différentiel linéaire de la forme

$$X' = AX + BU \quad , \quad \text{où}$$

- A est une matrice de taille n à coefficients constants,
- B est une matrice constante à n lignes et m colonnes,
- U est un vecteur dont les coefficients $u_1, u_2, \ldots, u_m$ sont des paramètres.

On dit que c'est un *système contrôlé, avec paramètre de contrôle* U. Comme dans le cadre des itérations affines (page 191), étudions si par une commande convenable, on peut amener la réponse $X(t)$ dans l'état qu'on veut.

2.1 Commandabilité

Définition
Le système contrôlé $X' = AX + BU$ est *commandable* si pour tout état initial X_0 et pour tout état final X_f, il existe un instant $t_f > 0$ et une fonction continue $U(t)$, $0 \leqslant t \leqslant t_f$, telle que la solution du système $X' = AX + BU(t)$ de condition initiale $X(0) = X_0$ vérifie $X(t_f) = X_f$.

Quand un système est commandable, il est possible d'atteindre, en un temps fini, n'importe quel objectif X_f à partir de n'importe quel état initial X_0, au moyen d'une commande $U(t)$ appropriée. Comme pour les itérations affines, introduisons la matrice de commandabilité.

Définition
La *matrice de commandabilité* du système est la matrice $C=[\,A^{n-1}B \;\; A^{n-2}B \;\; \ldots \;\; AB \;\; B\,]$ obtenue en juxtaposant les n matrices $A^{n-1}B, A^{n-2}B, \cdots, AB, B$. La matrice C possède n lignes et nm colonnes.

Critère de commandabilité. *Le système linéaire contrôlé $X' = AX + BU$ est commandable si et seulement si sa matrice de commandabilité est de rang n.*

Démonstration que la condition est suffisante. Choisissons comme résolvante du système $X' = AX$ la matrice $R = R_0$ telle que $R(0) = I_2$ (page 488). On a donc aussi $R(t+t') = R(t)R(t')$ et $R(t)^{-1} = R(-t)$. Remarquons que la matrice $B(^tB)$ est carrée de taille n et posons

$$M(s) = R(-s)B\,(^tB)\left[^tR(-s)\right] \quad \text{et} \quad Q = \int_0^{t_f} M(s)\,ds$$

l'intégration se faisant coefficient par coefficient. Si la matrice Q est inversible, définissons la commande

$$U(s) = (^tB)\left[^tR(-s)\right]Q^{-1}Y \ , \ \text{avec } Y = R(-t_f)X_f - X_0.$$

Une solution particulière du système $X' = AX + BU(t)$ est $S(t) = R(t)K(t)$, avec $K(t) = \int_0^t R(-s)BU(s)\,ds$ (page 492). On a $R(-s)BU(s) = R(-s)B(^tB)\left[^tR(-s)\right]Q^{-1}Y$. Comme l'intégrale est une opération linéaire, il vient

$$K(t_f) = \int_0^{t_f} R(-s)B(^tB)\left[^tR(-s)\right]Q^{-1}Y\,ds = \left[\int_0^{t_f} R(-s)B(^tB)\left[^tR(-s)\right]ds\right]Q^{-1}Y$$

c'est-à-dire $K(t_f) = QQ^{-1}Y = Y$, d'où $S(t_f) = R(t_f)Y$. Ajoutons à $S(t)$ la solution $R(t)X_0$ du système homogène. On obtient une solution $X(t)$ telle que $X(t_f) = R(t_f)X_0 + S(t_f)$, donc

$$X(t_f) = R(t_f)X_0 + R(t_f)\big[R(-t_f)X_f - X_0\big] = R(t_f)X_0 + X_f - R(t_f)X_0 = X_f$$

$$X(0) = R(0)X_0 + S(0) = X_0 \ , \quad \text{car } R(0) = I_n \text{ et } S(0) = 0.$$

La commande continue $U(t)$ permet donc d'atteindre l'état final X_f à partir de X_0.
Supposons maintenant que la matrice de commandabilité est de rang n. Faisons l'hypothèse que Q n'est pas inversible et cherchons une contradiction. Pour tout s, la matrice $M(s) = \big[R(-s)B\big]\,{}^t\big[R(-s)B\big]$ est symétrique et positive, donc Q aussi. Puisqu'on suppose Q non inversible, c'est qu'il existe un vecteur $V \in \mathbb{R}^n$ non nul, tel que ${}^tVQV = 0$ (page 218). Cela s'écrit encore $\int_0^{t_f} ({}^tV)M(s)V\,ds = \int_0^{t_f} \|({}^tV)R(-s)B\|^2\,ds = 0$, où $\|\ \|$ désigne la norme euclidienne. La fonction sous le signe intégrale est continue et positive ou nulle, son intégrale est nulle, donc elle est nulle (page 288). On a donc $({}^tV)R(s)B = 0$ pour tout $s \in [-t_f, 0]$. En dérivant, il vient $({}^tV)R'(s)B = ({}^tV)AR(s)B = 0$, car $R' = AR$, et en dérivant k fois, on obtient $({}^tV)A^kR(s)B = 0$ pour tout s entre $-t_f$ et 0. En particulier, en prenant $s = 0$, on a $R(0) = I_n$ et donc $({}^tV)A^kB = 0$. Cette égalité est vraie pour tout k, donc

$$({}^tV)\,[\,A^{n-1}B \ \ A^{n-2}B \ \ldots \ AB \ \ B\,] = [\,({}^tV)A^{n-1}B \ \ ({}^tV)A^{n-2}B \ \ldots \ ({}^tV)AB \ \ ({}^tV)B\,] = 0$$

ou encore $({}^tV)C = 0$ en appelant C la matrice de commandabilité. Comme le vecteur-ligne ${}^tV$ n'est pas nul, cela exprime que les n lignes de C ne sont pas indépendantes : c'est une contradiction. ■

Exemple. Pour le modèle de four présenté dans l'exemple précédent, la matrice de commandabilité est $C = [\,AB \ \ B\,] = \begin{bmatrix} -(a_er_e + a_ir_i)/c_p^2 & 1/c_p \\ a_ir_i/c_ic_p & 0 \end{bmatrix}$. Les deux lignes ne sont pas proportionnelles, donc C est de rang 2 et le système est commandable.

Ce résultat théorique n'est qu'un premier élément de réponse à la question posée, car dans la pratique, on ne dispose que d'une commande $u(t)$ à valeurs positives ou nulles.

2.2 Introduction au rétro-contrôle

Pour une condition initiale donnée, la réponse $X(t)$ d'un système différentiel contrôlé $X' = AX + BU$ dépend du vecteur de contrôle $U = (u_1, \ldots, u_m)$ dont les coefficients agissent par les formules linéaires définies par B. Pour réaliser une commande automatique, on fait à tout instant dépendre U de l'état X. On réalise ce couplage au moyen de différents dispositifs, comme par exemple un thermostat, qui mesurent l'état X et calculent la commande $U(X)$. On dit qu'on a réalisé un *bouclage* du système. Le système fonctionne alors en mode automatique.
Considérons seulement le cas d'un bouclage linéaire : le vecteur $U(X)$ est de la forme $-KX$, où K est une matrice à m lignes et n colonnes ; la matrice BK est alors carrée de taille n, comme la matrice A. Le système bouclé s'écrit $X' = AX - BKX$, ou encore

$$(c) \qquad X' = (A - BK)X$$

En général, on cherche à stabiliser dans le temps l'état $X(t)$: il s'agit donc de trouver une matrice K pour que le système différentiel (c) ait un équilibre stable. Lorsque

c'est possible, on dit que le système contrôlé $X' = AX + BU$ peut être *stabilisé par bouclage linéaire*.

Critère de régulation. *Tout système linéaire contrôlé commandable peut être stabilisé par bouclage linéaire.*

Nous verrons qu'on peut même trouver une matrice de bouclage K pour que les valeurs propres du système (c) soient des nombres $\lambda_1, \lambda_2, \ldots, \lambda_n$ donnés à l'avance : en choisissant $\lambda_i < 0$ pour tout i, on obtient un système (c) dont toutes les solutions tendent vers l'équilibre $X = 0$, avec des coordonnées en $e^{\lambda_i t}$ (si les λ_i sont deux à deux différents, toute solution est combinaison de solutions propres). Il est donc possible de régler la stabilisation à un niveau plus ou moins fort.
La démonstration du critère montre comment calculer effectivement une matrice de bouclage K.

Démonstration. Pour simplifier, supposons $n = 3$ et faisons la démonstration dans le cas d'une commande numérique : B est une matrice à une seule colonne et $U = u$ est un nombre ($m = 1$). Le système contrôlé est $X' = AX + Bu$ et dans sa matrice de commandabilité $C = [\,A^2B \;\, AB \;\, B\,]$, les produits A^iB sont des colonnes : la matrice C est donc carrée de taille 3. Supposons que le système est commandable, donc la matrice C est inversible. On a $I_3 = C^{-1}[\,A^2B \;\, AB \;\, B\,] = [\,C^{-1}A^2B \;\, C^{-1}AB \;\, C^{-1}B\,]$, donc les colonnes $C^{-1}A^2B$, $C^{-1}AB$ et $C^{-1}B$ sont les vecteurs canoniques $E_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $E_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ et $E_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Posons $L = [\,1 \;\, 0 \;\, 0\,]C^{-1} = ({}^tE_1)C^{-1}$. Puisque ${}^tE_1 = LC = [\,LA^2B \;\, LAB \;\, LB\,]$, on a

$$LA^2B = 1 \quad \text{et} \quad LAB = LB = 0.$$

Soit $P = \begin{bmatrix} L \\ LA \\ LA^2 \end{bmatrix}$. C'est une matrice carrée de taille 3 et l'on a $PC = \begin{bmatrix} LC \\ LAC \\ LA^2C \end{bmatrix} = \begin{bmatrix} 1 \;\, 0 \;\, 0 \\ LAC \\ LA^2C \end{bmatrix}$.
Montrons que la matrice PC est inversible.

- La deuxième ligne est $LAC = [\,LA^3B \;\, LA^2B \;\, LAB\,] = [\,LA^3B \;\, 1 \;\, 0\,]$. D'après le théorème de Cayley-Hamilton (page 152), on a la relation $A^3 = -pA^2 - qA - rI_3$, où $-(z^3 + pz^2 + qz + r)$ est le polynôme caractéristique de A. Il vient donc $LA^3B = -pLA^2B - qLAB - rLB = -p$ et $LAC = [\,-p \;\, 1 \;\, 0\,]$.
- De même, $LA^4B = -pLA^3B - qLA^2B - rLAB = p^2 - q$ et donc $LA^2C = [\,LA^4B \;\, LA^3B \;\, LA^2B\,] = [\,p^2 - q \;\, -p \;\, 1\,]$.

Les lignes de PC sont en échelon, donc indépendantes, la matrice PC est inversible et par suite aussi P. Dans le système $X' = AX + Bu$, faisons le changement d'inconnue $Y = PX$. Il vient $Y' = PX' = PAX + PBu$, ou encore

(1) $$Y' = (PAP^{-1})Y + PBu$$

Pour calculer la matrice de ce système, rappelons-nous que les lignes d'une matrice M sont les produits $({}^tE_i)M$. La première ligne de PAP^{-1} est donc

$$\begin{aligned} ({}^tE_1)PAP^{-1} &= LAP^{-1}\ , \quad \text{car } ({}^tE_1)P = L \text{ est la première ligne de } P \\ &= {}^tE_2\ , \quad \text{car } LA = ({}^tE_2)P \text{ est la deuxième ligne de } P. \end{aligned}$$

De même, la deuxième ligne de PAP^{-1} est $({}^tE_2)PAP^{-1} = LA^2P^{-1} = {}^tE_3$, car $({}^tE_2)P = LA$ et $LA^2 = ({}^tE_3)P$ est la troisième ligne de P. Enfin, on a

$({}^tE_3)PAP^{-1} = LA^3P^{-1} = -pLA^2P^{-1} - qLAP^{-1} - rLP^{-1} = -p{}^tE_3 - q{}^tE_2 - r{}^tE_1 = [-r \;\; -q \;\; -p]$

Ainsi $PAP^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -r & -q & -p \end{bmatrix}$. Puisque $PB = \begin{bmatrix} LB \\ LAB \\ LA^2B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, on obtient

$$(1') \qquad Y' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -r & -q & -p \end{bmatrix} Y + E_3\, u$$

Un bouclage linéaire pour $(1')$ est de la forme $u = -KY$, où $K = [-k_1 \;\; -k_2 \;\; -k_3]$ est une matrice-ligne formée de scalaires, et le système bouclé s'écrit $Y' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -r & -q & -p \end{bmatrix} - E_3KY$.

La matrice E_3K, carrée de taille 3, a ses deux premières lignes nulles et la troisième est K, donc le système bouclé est

$$(2) \qquad Y' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -r-k_1 & -q-k_2 & -p-k_3 \end{bmatrix} Y$$

Notons M la matrice du système différentiel (2). Le polynôme caractéristique de M est $Q_M = -\big(z^3 + (p+k_3)z^2 + (q+k_2)z + (r+k_1)\big)$. Il est possible de choisir k_1, k_2, k_3 de manière que les racines de Q_M soient des nombres donnés $\lambda_1, \lambda_2, \lambda_3$: en effet, il suffit d'écrire l'égalité $Q_M = -(z-\lambda_1)(z-\lambda_2)(z-\lambda_3)$, d'identifier les coefficients des puissances de z et d'en tirer k_1, k_2, k_3 en fonction des λ_i. En choisissant tous les λ_i strictement négatifs, on obtient un bouclage stabilisant pour (1).

Puisqu'on a seulement fait un changement linéaire d'inconnue, le système différentiel en X obtenu à partir de (2) est également stabilisé. Il s'écrit $X' = (A - BKP)X$, donc $u = -KPX$ est un bouclage stabilisant pour le système $X' = AX + Bu$. ■

Exemple : une régulation *in vivo*. Considérons deux populations d'insectes, l'une servant de proies pour l'autre. En milieu naturel, quand le nombre moyen de proies est x_m, le nombre de prédateurs s'établit à y_m. Quand on élève ces populations dans un vivarium, le nombre de proies est $x_m + x$ et le nombre de prédateurs est $y_m + y$. Supposons pour simplifier que, dans ce milieu artificiel, les équations d'évolution sont

$$(s) \qquad \begin{cases} \dot{x} = 2x - 3y \\ \dot{y} = 2x - y \end{cases}$$

La matrice $A = \begin{bmatrix} 2 & -3 \\ 2 & -1 \end{bmatrix}$ du système différentiel a pour polynôme caractéristique $\begin{vmatrix} 2-z & -3 \\ 2 & -1-z \end{vmatrix} = z^2 - z + 4$ et les valeurs propres de A sont $\frac{1}{2} \pm \boldsymbol{i}\frac{\sqrt{15}}{2}$. La partie réelle est positive, donc les solutions du système ne restent pas bornées quand t tend vers $+\infty$ (voir les trajectoires page 491) : on assiste à une croissance illimitée des deux populations.

Supposons que dans des limites raisonnables, la température dans le vivarium agit sur le taux d'éclosion des œufs de proies sans affecter les prédateurs, avec un contrôle

de la forme

$$\begin{cases} \dot{x} &= 2x - 3y + au \\ \dot{y} &= 2x - y \end{cases}$$

où u est l'écart à la température moyenne et a une constante positive. Peut-on réaliser un bouclage de ce système de manière à stabiliser les populations quels que soient les effectifs initiaux ?

Le système contrôlé est $\dot{X} = AX + \begin{bmatrix} a \\ 0 \end{bmatrix} u$ et un bouclage s'écrit $u = -kx$, d'où le système bouclé

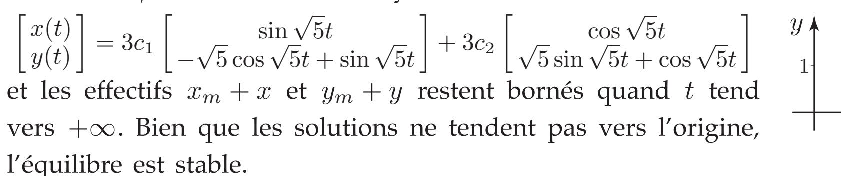
$$(b) \qquad \dot{X} = AX + \begin{bmatrix} -akx \\ 0 \end{bmatrix} = \begin{bmatrix} (2-ak)x & -3 \\ 2 & -1 \end{bmatrix} X$$

En posant $r = ak$, le polynôme caractéristique de la matrice de (b) est

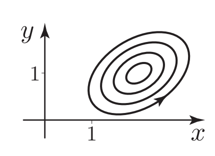
$$P = \begin{vmatrix} 2-r-z & -3 \\ 2 & -1-z \end{vmatrix} = z^2 + (r-1)z + r + 4$$

Pour que l'équilibre $(0, 0)$ soit stable, il faut et il suffit que les coefficients $r-1$ et $r+4$ soient positifs ou nuls, c'est-à-dire que l'on ait $r \geqslant 1$, ou encore $k \geqslant 1/a$.

- Pour $k = 1/a$, les solutions du système bouclé sont

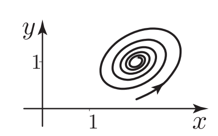


$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = 3c_1 \begin{bmatrix} \sin\sqrt{5}t \\ -\sqrt{5}\cos\sqrt{5}t + \sin\sqrt{5}t \end{bmatrix} + 3c_2 \begin{bmatrix} \cos\sqrt{5}t \\ \sqrt{5}\sin\sqrt{5}t + \cos\sqrt{5}t \end{bmatrix}$$



et les effectifs $x_m + x$ et $y_m + y$ restent bornés quand t tend vers $+\infty$. Bien que les solutions ne tendent pas vers l'origine, l'équilibre est stable.

- Pour $k > 1/a$, les valeurs propres sont de partie réelle strictement négative, donc les solutions tendent vers $(0,0)$: le système est stabilisé et les effectifs tendent vers leurs valeurs d'équilibre x_m et y_m d'autant plus vite que k est plus grand.



3. Systèmes différentiels généraux

On rencontre le plus souvent des systèmes différentiels où X' ne dépend pas linéairement de X.

Nous avons montré page 478 que l'on peut ramener l'étude d'un système différentiel quelconque à celle d'un système *autonome*, c'est-à-dire un système où la loi différentielle ne dépend pas du temps, comme par exemple $\begin{cases} x' &= f(x, y) \\ y' &= g(x, y) \end{cases}$. Dans la suite, nous ne considérerons donc que des systèmes autonomes.

On introduit le *vecteur d'état* $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ du système ; il décrit un domaine de $\mathbb{R}^n$.

Définition
Un *système différentiel (autonome)* s'écrit $X' = F(X)$, où $F(X) = \begin{bmatrix} f_1(X) \\ f_2(X) \\ \vdots \\ f_n(X) \end{bmatrix}$. Chacune des fonctions $f_1, \ldots, f_n$ prend des valeurs réelles et possède des dérivées partielles $\dfrac{\partial f_i}{\partial x_j}$ continues. Le *domaine*, ou *l'espace des états* du système différentiel, est le domaine de définition commun aux fonctions $f_1, \ldots, f_n$.

Une *solution* est une fonction $X(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}$ telle que $X'(t) = F\big(X(t)\big)$ pour tout t. Quand t varie, $X(t)$ décrit une courbe dans l'espace $\mathbb{R}^n$ des états. L'ensemble des points d'une courbe solution est une *trajectoire*.

Exemple. Une équation différentielle du second ordre de la forme $\ddot{x} = f(x, \dot{x})$ peut se transformer en un système différentiel. En posant $y = \dot{x}$, il vient en effet $\begin{cases} \dot{x} = y \\ \dot{y} = f(x,y) \end{cases}$. L'ensemble des états (x, y) du système s'appelle dans ce cas l'espace des phases de l'équation (voir page 462).

Propriétés des solutions

1) *Si X_0 est un point du domaine et t_0 un instant quelconque, il y a une unique solution $X(t)$ ayant pour condition initiale $X(t_0)=X_0$. La trajectoire associée s'appelle la trajectoire de X_0.*
2) *Si $X(t)$ est une solution, alors pour tout nombre a, la fonction $X(t+a)$ est solution.*

Nous avons démontré ces résultats pour les systèmes différentiels linéaires et pour les équations différentielles numériques autonomes ; ils sont encore vrais pour les systèmes différentiels vérifiant la définition ci-dessus. La propriété (2) est spécifique des systèmes autonomes : si $X(t)$ est solution de $X' = F(X)$ et si l'on pose $Y(t) = X(t+a)$, alors $Y'(t) = X'(t+a) = F\big(X(t+a)\big) = F\big(Y(t)\big)$ et $Y(t)$ est encore solution, de même trajectoire que $X(t)$. Pour étudier une solution, on peut donc supposer que sa condition initiale est $X(0) = X_0$.

Champ de vecteur associé au système. La fonction $X \mapsto F(X)$, définie sur le domaine du système différentiel, prend ses valeurs dans $\mathbb{R}^n$: il s'agit d'un champ de vecteurs (chapitre 14).

- Le champ de vecteurs $F(X)$ s'appelle le *champ associé* au système.
- Une trajectoire du système différentiel est tangente en chacun de ses points au vecteur du champ : une trajectoire est donc une ligne de champ.

3.1 Équilibres et trajectoires

Définition

Un point E du domaine est un *équilibre* du système différentiel si la fonction constante $X(t) = E$ est solution ; sa trajectoire est réduite au point E.

Pour qu'un point E soit un équilibre, il faut et il suffit que la fonction $X(t) = E$ vérifie $X'(t) = F\big(X(t)\big)$. Puisque $X'(t) = 0$, cette condition s'écrit $0 = F(E)$.

Un équilibre est un point E tel que $F(E) = 0$.

Propriétés des trajectoires

- *Deux trajectoires différentes ne se coupent pas.*
- *Les bouts d'une trajectoire sont en un point d'équilibre ou bien au bord du domaine.*
- *Si un bout de trajectoire est un équilibre E, les solutions sur cette trajectoire tendent vers E quand t tend vers $+\infty$ (ou quand t tend vers $-\infty$).*

Exemple : étude d'un système proies-prédateurs

Reprenons le système différentiel introduit page 477 :

$$\begin{cases} \dot{x} = ax - bxy \\ \dot{y} = -cy + dxy \end{cases}, \quad \text{où les constantes } a, b, c, d \text{ sont positives.}$$

Puisque x et y sont des effectifs de population, le domaine est le quart de plan formé des couples (x, y) tels que $x \geqslant 0$ et $y \geqslant 0$.

Le champ de vecteurs dans $\mathbb{R}^2$ associé au système est $F(x, y) = \begin{pmatrix} ax - bxy \\ -cy + dxy \end{pmatrix}$.

Recherche des équilibres. Les équilibres sont les solutions du système d'équations

$$\begin{cases} ax - bxy = 0 \\ -cy + dxy = 0 \end{cases} \iff \begin{cases} x(a - by) = 0 \\ y(-c + dx) = 0 \end{cases}$$

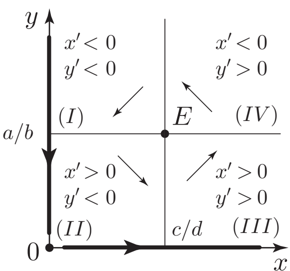
Si x ou y est nul, alors $x = y = 0$: il y a donc deux équilibres $O = (0,0)$ et $E = (c/d, a/b)$. Si les effectifs sont initialement $x_0 = c/d$ et $y_0 = a/b$, ils restent à ces valeurs d'équilibre.

Des solutions particulières

- En un point où $x = 0$, le vecteur du champ est $F(0, y) = (0, -cy)$; ces vecteurs étant verticaux, il y a des trajectoires portées par l'axe des ordonnées. Effectivement, pour $x = 0$, la seconde équation du système est $y' = -cy$, elle ne contient pas x et a pour solutions $y(t) = y_0 e^{-ct}$. Ainsi $X(t) = \big(0, y_0 e^{-ct}\big)$ est une solution du système. La trajectoire est l'axe des ordonnées positives parcouru dans le sens décroissant et $X(t)$ tend vers l'équilibre O quand t tend vers $+\infty$.
- De même, en tout point $(x, 0)$, le champ $F(x, 0) = (ax, 0)$ est horizontal, donc il y a des trajectoires portées par l'axe des abscisses. Pour $y = 0$, la première équation est $x' = ax$, donc $X(t) = (x_0 e^{at}, 0)$ est solution du système. La trajectoire est l'axe

des abscisses positives parcouru dans le sens croissant et $X(t)$ tend vers l'infini quand t tend vers $+\infty$.

Sens de variation des solutions



En chaque point (x, y) du domaine, l'orientation du vecteur $F(x, y)$ est fixée par le signe de ses composantes $\big[x(a-by), y(-c+dx)\big]$. Si par exemple on a $x > c/d$ et $y > a/b$, alors $a-by < 0$ et $-c+dx > 0$, donc le champ est dirigé « vers le haut à gauche » : en ce point, on a $x'(t) < 0$ et $y'(t) > 0$. Ainsi on obtient la représentation ci-contre qui permet de visualiser les variations de $x(t)$ et de $y(t)$ dans chacune des régions (I) à (IV).
Prenons une solution $X(t)$ dont la condition initiale est un point X_0 situé dans la région (I). Sa trajectoire T ne peut pas couper l'axe des ordonnées positives qui est lui-même une trajectoire, donc la courbe $X(t)$ pénètre dans la région (II). Dès lors, T ne peut pas couper l'axe des abscisses positives qui est une trajectoire, donc $X(t)$ pénètre dans la région (III). Ensuite, $X(t)$ pourrait éventuellement rester dans cette région (avec une asymptote horizontale située sous l'ordonnée a/b), mais nous allons voir que ce n'est pas le cas.

Équation des trajectoires. Cherchons y comme fonction de x. Puisqu'on a

$$y'(x) = \frac{\dot{y}}{\dot{x}} = \frac{y(-c+dx)}{x(a-by)}$$

il vient

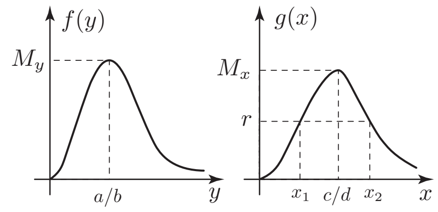
$$\frac{a-by}{y} y'(x) = \frac{-c+dx}{x}$$

qui est une équation à variables séparées. En intégrant chaque membre, on trouve $a\ln y - by + c\ln x - dx = k$, où k est une constante, et en prenant l'exponentielle, on obtient

(1) $$(y^a e^{-by})\,(x^c e^{-dx}) = K \ , \quad \text{avec } K > 0.$$

Les trajectoires sont les courbes d'équation (1), pour les différentes valeurs de K.

Les graphes des fonctions $f(y) = y^a e^{-by}$ et $g(x) = x^c e^{-dx}$ sont représentés ci-contre.

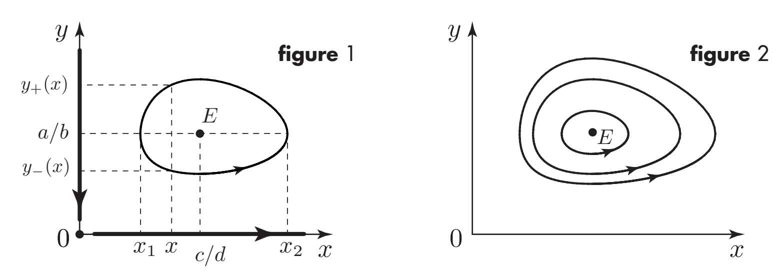


On a $f'(y) = y^{a-1}e^{-by}(a-by)$, le maximum de $f(y)$ est atteint en $y = a/b$ et vaut $M_y = (a/b)^a e^{-a}$; de même, le maximum de $g(x)$ est obtenu en $x = c/d$ et vaut $M_x = (c/d)^c e^{-c}$. Si $K > M_x M_y$, l'équation (1) n'a donc pas de solution. Si $K = M_x M_y$, la seule solution est l'équilibre $(c/d, a/b)$.
Supposons maintenant $K = rM_y$, où $0 < r < M_x$. L'équation $g(x) = r$ a deux solutions x_1 et x_2, avec $0 < x_1 < c/d < x_2$, et l'on a $r/g(x) > 1$ si x est en dehors du segment $[x_1, x_2]$. On en déduit que, pour x fixé, l'équation

(e) $$f(y) = \frac{r}{g(x)} M_y$$

n'a pas de solution en y lorsque x est hors de $[x_1, x_2]$, en a une seule $y = a/b$ lorsque $x = x_1$ ou $x = x_2$ et que si $x_1 < x < x_2$, alors (e) a deux solutions $y_-(x) < a/b < y_+(x)$. De plus, quand x tend vers x_1 ou x_2, $y_-(x)$ et $y_+(x)$ tendent vers a/b. Cela montre que les courbes définies par (1) sont fermées, comme sur la figure 1 ci-dessous.

Ainsi, dans le domaine $x > 0, y > 0$, les trajectoires du systèmes sont fermées : les effectifs oscillent autour de l'équilibre E (figure 2).



3.2 Stabilité d'un équilibre

Dans l'étude des systèmes différentiels linéaires, nous avons rencontré des équilibres stables et des équilibres instables.

Définition
Un point d'équilibre E est *stable* si pour tout $\varepsilon > 0$, la distance de $X(t)$ à E reste inférieure à ε quel que soit $t \geqslant 0$, pour toute solution $X(t)$ de condition initiale $X_0 = X(0)$ assez proche de E. Si de plus chaque $X(t)$ tend vers E quand t tend vers $+\infty$, on dit que E est un équilibre *asymptotiquement stable*.

Dans l'exemple proies-prédateurs du précédent paragraphe, l'équilibre E est stable, mais pas asymptotiquement stable.

Cas d'un système linéaire. Pour un système linéaire $X' = AX$, rappelons (page 490) que l'origine est un équilibre

- stable si A est diagonalisable avec toutes ses valeurs propres de partie réelle négative ou nulle,
- asymptotiquement stable si toutes les valeurs propres de A ont leur partie réelle strictement négative,
- instable si l'une au moins des valeurs propres de A est de partie réelle strictement positive.

3.3 Linéarisation autour d'un équilibre

Supposons que E est un point d'équilibre du système différentiel

$$(1) \qquad X' = F(X)$$

Pour approcher la fonction $F(X) = \big(f_1(X), \dots, f_n(X)\big)$ au voisinage de E, formons sa matrice jacobienne (page 367) au point E :

$$J(E) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

où les dérivées partielles sont calculées au point E. L'approximation affine de F au point $E = (e_1, \dots e_n)$ s'écrit (page 364)

$$F(E) + J(E) \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

avec $u_i = x_i - e_i$. Puisqu'on a supposé $F(E) = 0$, il vient simplement

$$J(E)U, \text{ en posant } U = X - E.$$

Puisque $U' = X'$, introduisons le système différentiel linéaire

$$(L) \qquad U' = J(E)U$$

Le système différentiel linéaire $U' = J(E)U$ s'appelle le *linéarisé* de (1) au point E. Puisque (L) est une approximation de (1) autour de E, on peut espérer que dans une région assez petite autour de E, les solutions de (1) ressemblent à celles de (L) autour de l'origine. C'est effectivement ce qui se passe assez généralement. Supposons désormais que la fonction F possède des dérivées partielles d'ordre aussi grand qu'on veut.

Définition

Le système (1) est *linéarisable autour de* E si les trajectoires de (1) ont, au voisinage de E, la même forme que celles de (L) autour de l'origine, le sens de parcours étant conservé ainsi que les directions dans lesquelles les trajectoires s'approchent de E.

Plus précisément, (1) est linéarisable s'il existe un changement de référentiel φ (page 24) qui transforme le point E en l'origine O et les trajectoires de (1) en celles de (L), avec la condition supplémentaire qu'au voisinage de E, on a $\varphi(X) = U + o(U)$, avec $\|o(U)\| \underset{U \to 0}{\ll} \|U\|$.

Conditions de linéarisation. *Si le système différentiel (1) vérifie l'une au moins des conditions suivantes, il est linéarisable autour de E.*

a) *Le système est de taille 2 et aucune valeur propre de la matrice $J(E)$ n'est de partie réelle nulle.*

b) *Les valeurs propres de $J(E)$ sont toutes de parties réelles strictement négatives, ou bien toutes de parties réelles strictement positives.*

c) *Il n'y a entre les valeurs propres $\lambda_1, \ldots, \lambda_n$ de $J(E)$ aucune relation de la forme $\lambda_j = \sum_{i=1}^{n} n_i \lambda_i$, avec des entiers n_i positifs ou nuls de somme au moins égale à 2.*

La condition (c) est dite « de non résonance ». Si l'on y remplace les λ_i par leur exponentielle e^{λ_i}, on obtient la condition de non résonance, rencontrée page 391, qui permet de linéariser une transformation ; les deux problèmes sont intimement liés.

Si le système (1) est linéarisable autour de E, cet équilibre a évidemment les mêmes propriétés de stabilité que l'équilibre O du système linéaire (L). Pour étudier la stabilité d'un équilibre E, on peut donc linéariser le système en ce point.

Supposons le système (1) linéarisable autour de E. Alors

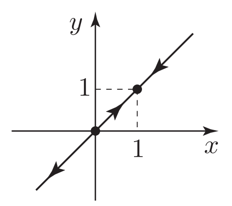
- *l'équilibre E est stable si et seulement si l'origine est stable pour le système linéarisé ;*
- *l'équilibre E est asymptotiquement stable si et seulement si l'origine est asymptotiquement stable pour le système linéarisé.*

Exemple. Étudions le système différentiel (1) $\begin{cases} x' = -x + 2y - x^2 \\ y' = x(1 + x - 2y) \end{cases}$

Recherche des équilibres du système. Ce sont les solutions du système d'équations $\begin{bmatrix} -x+2y-x^2 = 0 \\ x(1+x-2y) = 0 \end{bmatrix}$. On trouve $x = y = 0$ et si $x \neq 0$, il vient $1+x = 2y$, donc $1-x^2 = 0$, d'où $x = 1, y = 1$ ou bien $x = -1, y = 0$. Il y a trois équilibres : $O = (0,0)$, $A = (1,1)$ et $B = (-1,0)$.

Des trajectoires particulières. En tout point (x,y), le champ de vecteurs est $F(x,y) = \begin{bmatrix} -x+2y-x^2 \\ x(1+x-2y) \end{bmatrix}$. En un point de la droite d'équation $y = x$, le champ $F(x,x) = \begin{bmatrix} x-x^2 \\ x(1-x) \end{bmatrix}$ est dans la direction de cette droite, donc la droite $y = x$ est réunion de trajectoires. En restreignant le système à cette droite, on obtient l'équation différentielle $x' = \dfrac{dx}{dt} = x-x^2$.

Nous avons déjà étudié l'équation $x' = x-x^2$ dans l'exemple page 448 : sur une droite représentant l'espace des x, il y a cinq trajectoires : les deux équilibres $x = 0$ et $x = 1$, l'intervalle $]-\infty, 0[$ parcouru en décroissant (car $x' < 0$ pour $x < 0$), l'intervalle $]0,1[$ parcouru en croissant (car $x' > 0$ pour $0 < x < 1$) et l'intervalle $]1, +\infty[$ parcouru en décroissant (car $x' < 0$ pour $x > 1$).



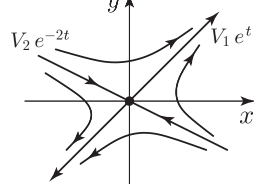
La figure ci-contre montre les trajectoires correspondantes pour le système (1). On voit déjà que l'origine n'est pas un équilibre stable.

Étude au voisinage d'un équilibre Nous allons linéariser le système en chacun des points d'équilibre. Pour former la matrice jacobienne $J(x,y)$ en un point quelconque, calculons les dérivées partielles de F :

$$\frac{\partial}{\partial x}(-x+2y-x^2) = -1-2x \quad , \quad \frac{\partial}{\partial y}(-x+2y-x^2) = 2$$
$$\frac{\partial}{\partial x}(x+x^2-2xy) = 1+2x-2y \quad , \quad \frac{\partial}{\partial y}(x+x^2-2xy) = -2x$$

Au point O. La matrice jacobienne de F en $(0,0)$ est $J = \begin{bmatrix} -1 & 2 \\ 1 & 0 \end{bmatrix}$. Son polynôme caractéristique est $(-1-z)(-z)-2 = (z-1)(z+2)$, donc les valeurs propres sont 1 et -2. Les vecteurs propres sont respectivement $V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ et $V_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$.

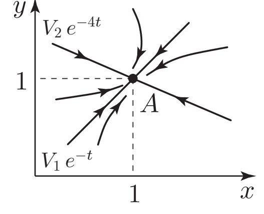
Pour avoir l'allure des trajectoires du système linéaire $U' = JU$, il suffit de se reporter à la figure 2 page 491.



Il y a deux trajectoires rectilignes dans la direction V_1, deux dans la direction V_2 et les autres sont de type « hyperbolique », avec pour asymptotes les directions propres V_1 et V_2. En linéarisant, on sait que, pour le système (1), l'allure des trajectoires autour de l'origine est la même.

Au point A. Au point $A = (1,1)$, la matrice jacobienne est $J = \begin{bmatrix} -3 & 2 \\ 1 & -2 \end{bmatrix}$. Le polynôme caractéristique est

$$(-3-z)(-2-z)-2 = (z+1)(z+4)\,,$$

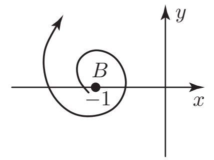


et les valeurs propres -1 et -4 ont pour vecteurs propres à nouveau $W_1 = V_1$ et $W_2 = V_2$.

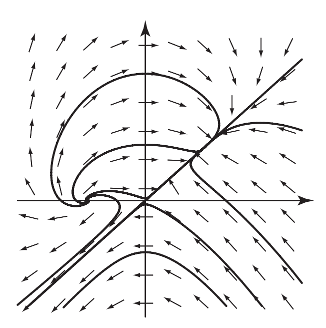
Mis à part les deux trajectoires rectilignes dans la direction V_2, les trajectoires du système linéaire $U' = JU$ sont tangentes à la direction propre de plus petite valeur propre (en valeur absolue), c'est-à-dire à W_1 (figure 1 page 491). Par linéarisation, on en déduit qu'au voisinage de A, les trajectoires de (1) ont la même allure. Le point A est un équilibre asymptotiquement stable.

Au point B. La matrice jacobienne en $B = (-1,0)$ est $J = \begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix}$. Le polynôme caractéristique est $(1-z)(2-z)+2$ et les valeurs propres sont $\lambda = (1/2)(3+\boldsymbol{i}\sqrt{7})$ et $\mu = (1/2)(3-\boldsymbol{i}\sqrt{7})$.

Les trajectoires du système $U' = JU$ sont des spirales qui s'écartent de l'équilibre, car la partie réelle des valeurs propres est strictement positive (figure 3 page 491). Il s'ensuit qu'au voisinage de B, les trajectoires de (1) sont également en forme de spirales qui s'écartent de B. Le point B n'est donc pas un équilibre stable. Pour trouver le sens de parcours, remarquons qu'en un point de l'axe des x, on a $y = 0$, donc $y' = x(1+x) < 0$ pour $-1 < x < 0$: les spirales sont donc parcourues dans le sens inverse des aiguilles d'une montre.



La figure ci-dessous montre le champ de vecteurs et quelques trajectoires du système.



3.4 Fonction de Liapounov

Soit E un point d'équilibre pour le système différentiel $X' = F(X)$.

Définition

Une *fonction de Liapounov* pour l'équilibre E est une fonction V à valeurs réelles, définie au voisinage de E, à dérivées partielles continues et satisfaisant les conditions suivantes :

i) $V(E) = 0$,

ii) $V(X) > 0$ pour tout $X \neq E$,

iii) le produit scalaire $\mathrm{Grad}_V(X) \cdot F(X)$ garde un signe constant.

Les conditions (i) et (ii) signifient que $V(X)$ présente un minimum au point E. Au voisinage de ce point, les lignes de niveau de V « encerclent » E et l'on sait que le vecteur gradient de V est orthogonal à ces lignes de niveau et dirigé vers l'extérieur :

- si $\mathrm{Grad}_V(X) \cdot F(X)$ est toujours négatif, alors $F(X)$ pointe vers l'intérieur des lignes de niveau de V : le long d'une trajectoire, le niveau de V diminue (figure 1) ;
- si au contraire $\mathrm{Grad}_V(X) \cdot F(X)$ est positif, $F(X)$ pointe vers l'extérieur des lignes de niveau et quand on parcourt une trajectoire, le niveau de V augmente (figure 2).

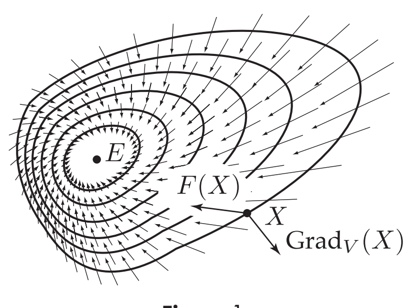



Figure 1

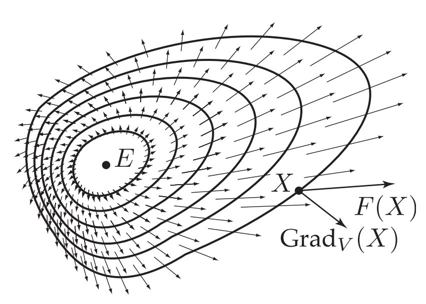


Figure 2

Soit $X(t)$ une solution de point initial $X_0 = X(0)$ assez proche de E. D'après la règle de dérivation d'une fonction composée, on a

$$\frac{d}{dt} V\big(X(t)\big) = \mathrm{Grad}_V\big(X(t)\big) \cdot X'(t) = \mathrm{Grad}_V\big(X(t)\big) \cdot F\big(X(t)\big)$$

Si le produit scalaire $\mathrm{Grad}_V(X) \cdot F(X)$ est toujours négatif ou nul, le niveau $V\big(X(t)\big)$ est décroissant le long de la trajectoire de X. On en déduit que pour $t \geqslant 0$, $X(t)$ est confinée à l'intérieur de la ligne du niveau initial : l'équilibre est stable.
Montrons que si $\mathrm{Grad}_V(X) \cdot F(X)$ est strictement négatif, alors $X(t)$ tend vers E.

> La fonction niveau $n(t) = V\big(X(t)\big)$ est strictement décroissante, donc quand t tend vers $+\infty$, $n(t)$ a une limite m positive ou nulle. Puisque E est, d'après (i) et (ii), le seul point de niveau nul, il faut montrer que $m = 0$. Raisonnons par l'absurde en supposant $m > 0$. Dans ce cas, la solution resterait dans la « couronne » formée des points de niveau compris entre m et le niveau initial ; cela implique que $X(t)$ est défini pour tout $t \geqslant 0$. Or la couronne étant compacte, la fonction continue $\frac{dn}{dt} = \mathrm{Grad}_V\big(X(t)\big) \cdot F\big(X(t)\big)$ a un maximum $c < 0$ dans cette couronne (page 362). En intégrant, on obtiendrait $n(t) \leqslant n(0) + ct$ pour $t \geqslant 0$, et en prenant t assez grand, il viendrait $n(t) < 0$, ce qui est absurde.

Énonçons ces résultats.

Critère de stabilité de Liapounov. *Supposons que V est une fonction de Liapounov pour l'équilibre E.*

- *Si $\mathrm{Grad}_V(X) \cdot F(X) \leqslant 0$ pour tout X, l'équilibre est stable.*
- *Si $\mathrm{Grad}_V(X) \cdot F(X) < 0$ pour tout $X \neq E$, l'équilibre est asymptotiquement stable.*

Les lignes de niveau d'une fonction de Liapounov permettent de décrire le bassin de stabilité ou de stabilité asymptotique de l'équilibre, c'est-à-dire l'ensemble des points initiaux à partir desquels la solution va rester proche de E ou tendre vers E.

Exemple. Prenons le système différentiel du pendule sur lequel on agit par une force $g(\dot{x})\dot{x}$ fonction de sa vitesse, où l'on suppose g dérivable. En posant $y = \dot{x}$, le système devient

$$\begin{cases} \dot{x} &= y \\ \dot{y} &= -\sin x - g(y)y \end{cases}$$

L'origine $O = (0, 0)$ est un équilibre. La dérivée de $g(y)y$ en $y = 0$ est $g(0)$, donc le système linéarisé en O a pour matrice $A = \begin{bmatrix} 0 & 1 \\ -1 & -g(0) \end{bmatrix}$. Le polynôme caractéristique est $z^2 + g(0)z + 1$.

Supposons $g(0) > 0$. Le polynôme caractéristique de A a ses coefficients positifs et la somme des racines est $-2g(0) < 0$, donc les racines sont de partie réelle strictement négative (page 490). L'équilibre O est asymptotiquement stable pour le système linéarisé et, par linéarisation, on en déduit que l'origine est aussi asymp-

totiquement stable pour le système initial. Pour de petites vitesses, $-g(\dot{x})\dot{x}$ a le signe opposé de $\dot{x}$, donc le pendule est amorti.

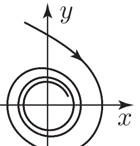
Supposons $g(0) = 0$ et $g(y) \geqslant 0$ pour tout y. Cette fois, le système n'est plus linéarisable autour de l'origine, car les valeurs propres du linéarisé sont $\pm\, \boldsymbol{i}$. Mais prenons la fonction énergie $V(x, y) = (1/2)y^2 + 1 - \cos x$. Puisque $1 - \cos x \geqslant 0$, $V(x, y)$ est positif ou nul et n'est nul que si $y = 1 - \cos x = 0$: dans le domaine $-\pi < x \leqslant \pi$, cela implique $y = x = 0$. On a

$$\mathrm{Grad}_V(X) = \begin{bmatrix} \dfrac{\partial V}{\partial x} & \dfrac{\partial V}{\partial y} \end{bmatrix} = [\sin x \quad y], \text{ donc}$$

$$\mathrm{Grad}_V(X) \cdot F(X) = [\sin x \quad y] \begin{bmatrix} y \\ -\sin x - g(y)y \end{bmatrix} = -y^2 g(y)$$

Ce produit scalaire est toujours négatif ou nul, donc l'équilibre est stable.

- Si $g(y)$ est n'est pas identiquement nul, l'énergie $V\big(X(t)\big)$ diminue le long d'une trajectoire : il n'y a pas de trajectoire fermée et donc pas de solution périodique.
- Si l'on a $g(y) > 0$ pour $y \neq 0$, alors $\mathrm{Grad}_V(X) \cdot F(X)$ est strictement négatif sauf aux points où $y = 0$ et cela implique que les solutions tendent vers 0 : l'origine est un équilibre asymptotiquement stable. Le dessin ci-contre montre une trajectoire dans le cas $g(y) = y^2$.



- Si $g(y)$ est identiquement nul, le système est celui du pendule simple et sur chaque trajectoire, le niveau de V reste constant. Les trajectoires proches de l'origine sont fermées (figure 2 page 463) et l'équilibre n'est pas asymptotiquement stable.

Remarque

Souvent, on cherche une fonction de Liapounov en s'inspirant des caractéristiques physiques du système, comme dans l'exemple ci-dessus.
Voici un autre moyen de construire des fonctions ayant les propriétés (i) et (ii) de la définition. Considérons une matrice S symétrique définie positive (page 218). Si l'on pose $V(X) = ({}^tX)SX$ pour tout vecteur $X \in \mathbb{R}^n$, alors on a $V(0) = 0$ et $V(X) > 0$ pour tout $X \neq 0$.

3.5 Systèmes hamiltoniens

En Mécanique, un mouvement $X(t)$ dans un champ de force qui dérive d'un potentiel $-V$ est solution d'un système différentiel du second ordre, du type

$$(1) \qquad \ddot{X} = -\mathrm{Grad}_V(X)$$

En plus de la variable $X = (x_1, \ldots, x_n)$, introduisons les variables $p_i = \dot{x}_i$ et posons $P = (p_1, \ldots, p_n)$. L'espace des variables $(x_1, x_2, \ldots, x_n, p_1\ p_2, \ldots, p_n) = (X, P)$ s'appelle *l'espace des phases* ; il est de dimension $2n$. Comme dans l'étude du pendule page 462, définissons

- l'énergie cinétique $T(P) = \frac{1}{2}(p_1^2 + \cdots + p_n^2)$
- et l'énergie totale $H(X,P) = T(P) + V(X)$, somme de l'énergie cinétique et de l'énergie potentielle.

On a $\dfrac{\partial H}{\partial p_i} = \dfrac{\partial T}{\partial p_i} = p_i$ et $\dfrac{\partial H}{\partial x_i} = \dfrac{\partial V}{\partial x_i}$. La i-ième équation de (1) est $\dot{p}_i = \ddot{x}_i = -\dfrac{\partial V}{\partial x_i}$, donc (1) est équivalent à

$$(2) \qquad \begin{cases} \dot{x}_i &= \dfrac{\partial H}{\partial p_i} \\ \dot{p}_i &= -\dfrac{\partial H}{\partial x_i} \end{cases}$$

La fonction H s'appelle le *Hamiltonien* du système (1). Puisque le Hamiltonien est l'énergie totale d'un état du système, cette fonction prend une valeur constante le long de toute solution de (1) (page 425).

Si $X(t)$ est une solution de (1), la fonction $H\big(X(t)\big)$ est constante.

On vérifie en effet que l'on a $\dfrac{d}{dt}H\big(X(t)\big) = \sum_{i=1}^n \dfrac{\partial H}{\partial x_i}\dot{x}_i + \sum_{i=1}^n \dfrac{\partial H}{\partial p_i}\dot{p}_i = 0$, car d'après (2), $\dfrac{\partial H}{\partial x_i}\dot{x}_i + \dfrac{\partial H}{\partial p_i}\dot{p}_i = 0$ pour tout i.

Ce résultat signifie que dans l'espace des phases, chaque trajectoire de (2) est incluse dans une ligne, ou en général une surface, d'équation $H(X,P) =$ constante, c'est-à-dire dans une surface de niveau de la fonction H.
Pour un tel système de taille deux, on peut ainsi calculer l'équation des trajectoires.

Exemple. Soit le système différentiel $\begin{cases} x' &= 4y^3+x \\ y' &= -(4x^3+y) \end{cases}$. Posons $f(x,y) = 4y^3+x$ et $g(x,y) = 4x^3+y$. On a $\dfrac{\partial g}{\partial y} = \dfrac{\partial f}{\partial x} = 1$, donc il existe une fonction H telle que $\left[\dfrac{\partial H}{\partial x} \;\; \dfrac{\partial H}{\partial y}\right] = [\,g(x,y) \;\; f(x,y)\,]$ (page 420). Le système s'écrit

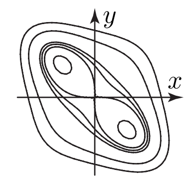
$$\begin{cases} x' &= f(x,y) = \dfrac{\partial H}{\partial y}(x,y) \\ y' &= -g(x,y) = -\dfrac{\partial H}{\partial x}(x,y) \end{cases}$$

et d'après ce qui précède, on en déduit que les trajectoires ont pour équation $H(x,y) = K$, où K est constant. Calculons $H(x,y)$.

$$\frac{\partial H}{\partial y} = 4y^3+x \quad \text{donc } H(x,y) = y^4+xy+u(x)$$

$$\frac{\partial H}{\partial x} = y+u'(x) = 4x^3+y \quad \text{donc } u'(x) = 4x^3 \text{ et } u(x) = x^4+ \text{ constante}$$

On trouve ainsi $H(x,y) = x^4+y^4+xy$ et les trajectoires du système, représentées ci-contre, ont pour équation $x^4+y^4+xy=$ constante. Les équilibres sont donnés par $f(x,y) = g(x,y) = 0$; il en a trois : $(0,0)$, $(1/2,-1/2)$ et $(-1/2,1/2)$. Les deux derniers sont visiblement stables.



4. Dépendance par rapport à la condition initiale

On est parfois intéressé par la manière dont une solution d'un système différentiel $X' = F(X)$ dépend de sa condition initiale. Pour tout vecteur $a = [a_1, \dots, a_n]$, notons $X(t, a)$ la solution de condition initiale $X(0, a) = a$, et posons

$$Y_i(t, a) = \frac{\partial}{\partial a_i} X(t, a)$$

Puisque $X(0, a) = a$ pour tout a, on a $Y_i(0, a) = E_i$, le i-ème vecteur canonique. Par définition $\frac{\partial}{\partial t} X(t, a) = F\big(X(t, a)\big)$, donc en dérivant par rapport à a_i, il vient

$$\frac{\partial}{\partial a_i} \frac{\partial}{\partial t} X(t, a) = J_F\big(X(t, a)\big) \frac{\partial}{\partial a_i} X(t, a) = J_F\big(X(t, a)\big) Y_i(t, a)$$

où J_F est la matrice jacobienne de F. Puisqu'on peut dériver dans l'ordre qu'on veut, on en déduit

$$\text{(1)} \qquad \frac{\partial}{\partial t} Y_i(t, a) = J_F\big(X(t, a)\big) Y_i(t, a)$$

La fonction $t \mapsto Y_i(t, a)$ est la solution du système différentiel linéaire (1) telle que $Y_i(0) = E_i$.

En général, ce système n'est pas à coefficients constants, car la matrice $J_F\big(X(t, a)\big)$ dépend de t.

Remarque

Lorsque F a des dérivées partielles continues, les solutions du système différentiel $X' = F(X)$ ont ainsi des dérivées partielles par rapport aux coordonnées de la condition initiale : les solutions dépendent donc continûment de leur condition initiale.

Cas où a est un équilibre. Supposons que la condition initiale a est un point d'équilibre du système. Dans ce cas, on a $X(t, a) = a$ pour tout t et le système différentiel (1) devient simplement

$$\frac{\partial}{\partial t} Y_i(t, a) = J_F(a) Y_i(t, a)$$

C'est un système différentiel linéaire à coefficients constants, de matrice $J_F(a)$. En calculant la solution $Y_i(t)$ telle que $Y_i(0) = E_i$, on trouve que pour un petit accroissement δa_i de la i-ème coordonnée de a, on a

$$X\big(t, (a_1, \dots, a_i + \delta a_i, \dots, a_n)\big) \simeq a + (\delta a_i) Y_i(t, E)$$

5. Un exemple de prévision en épidémiologie

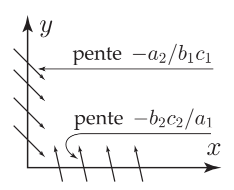
Voici un exemple de modélisation pour une épidémie à durée d'incubation courte par rapport au temps pendant lequel un sujet est contagieux ; ce modèle a été proposé initialement pour certains types de MST. On suppose la maladie non mortelle, la propagation se faisant dans un groupe constitué de c_1 hommes et de c_2 femmes. On considère

- des hommes infectés, soignés dans une proportion a_1 ;
- des femmes infectées, soignées dans une proportion a_2 : ces cas étant assez souvent asymptomatiques, a_2 est petit devant a_1 ;
- un taux d'infection masculine proportionnel au nombre d'hommes sains et de femmes infectées (coefficient b_1) ;
- un taux d'infection féminine proportionnel au nombre de femmes saines et d'hommes infectés (coefficient b_2).

Notons $x(t)$ le nombre d'hommes infectés et $y(t)$ le nombre de femmes infectées. Il y a donc $c_1 - x(t)$ hommes et $c_2 - y(t)$ femmes susceptibles de tomber malades. Les équations s'écrivent

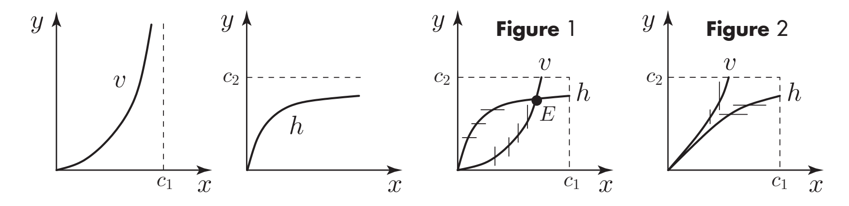
$$(1) \qquad \begin{cases} \dot{x} = -a_1 x + b_1(c_1 - x)y \\ \dot{y} = -a_2 y + b_2(c_2 - y)x \end{cases}, \quad \text{où } 0 < x(0) < c_1 \text{ et } 0 < y(0) < c_2.$$

Étudions le champ de vecteurs $F(x, y)$ associé.



- En un point de l'axe des $y > 0$, on a $F(0, y) = (b_1 c_1 y\, , -a_2 y)$, donc le champ pointe « en bas à droite ». En un point de l'axe des $x > 0$, on a $F(x, 0) = (-a_1 x\, , b_2 c_2 x)$ et le champ pointe « en haut à gauche ».
- Cherchons les points (x,y) où le champ est vertical : on doit avoir $-a_1x+b_1(c_1-x)y = 0$, d'où $y = \dfrac{a_1 x}{b_1(c_1 - x)} = v(x)$.
- Les points où le champ est horizontal vérifient $-a_2y + b_2(c_2-y)x = 0$, d'où $y = \dfrac{b_2 c_2 x}{a_2 + b_2 x} = h(x)$.

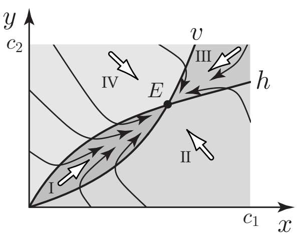
Voici les graphes des fonctions $v(x)$ et $h(x)$: selon la valeur des constantes, il y a deux dispositions possibles pour ces courbes :



La figure 1 montre le cas $a_1 a_2 < b_1 b_2 c_1 c_2$. Au point E d'intersection des courbes, on a $v(x) = h(x) = y$, donc le champ est nul : E est un point d'équilibre.
La figure 2 correspond au cas $a_1 a_2 > b_1 b_2 c_1 c_2$.

Cas $a_1 a_2 < b_1 b_2 c_1 c_2$

Il y a quatre régions et la figure ci-contre montre dans chacune d'elles la direction du champ. Par exemple, dans la région (I), on a $v(x) < y < h(x)$, donc $\dot{x} > 0$ et $\dot{y} > 0$; dans la région (II), on a $y < h(x)$ et $y < v(x)$, donc $\dot{x} < 0$ et $\dot{y} > 0$. On en déduit le sens de variation des trajectoires.



- Considérons une solution qui a un point dans la région (I) à un certain instant t_0. Comme le champ est horizontal le long du bord supérieur de (I) et que la solution progresse dans le sens x et y croissants, la solution ne peut pas s'échapper de (I) par en haut ; de même, le champ étant vertical le long du bord inférieur, la solution ne peut s'échapper par en bas. Ainsi, la solution reste dans (I) pour tout $t \geqslant t_0$:

  la région (I) est une *zone piège*.

  On en déduit qu'une solution qui a pénétré dans (I) tend vers l'équilibre E quand t tend vers $+\infty$.
- De même, la région (III) est une zone piège : toute solution qui a pénétré dans (III) y reste ultérieurement et donc tend vers E.
- Considérons une solution $X(t) = \big(x(t), y(t)\big)$ de point initial dans (II). Tant que $X(t)$ est dans (II), x décroît et y croit : $X(t)$ doit donc pénétrer dans l'une des régions (I) ou (III) et dès lors, $X(t)$ tend vers E quand t tend vers l'infini. La seule exception concerne le cas où $X(t)$ tend directement vers E en restant dans la région (II).
- De même, une solution initialement dans (IV) doit pénétrer dans (I) ou dans (III), à moins qu'elle ne tende vers E en restant dans (IV) ; en tous cas, cette solution tend vers E.

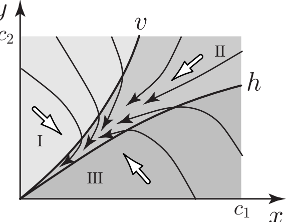
Ainsi toutes les solutions tendent vers E quand t tend vers l'infini. L'équilibre E est asymptotiquement stable.

Pour calculer les coordonnées α, β de E, on résout l'équation $v(x) = h(x)$ et l'on a $\beta = h(\alpha) = v(\alpha)$.

Conclusion : Si $\dfrac{b_1 c_1}{a_2} \dfrac{b_2 c_2}{a_1} > 1$, le nombre d'hommes infectés tend vers α et le nombre de femmes infectées vers β. L'épidémie prend un régime stationnaire.

Cas $a_1 a_2 > b_1 b_2 c_1 c_2$

Il y a trois régions et dans chacune d'elles, le champ a la direction indiquée ci-contre. Par exemple, dans la région (II), on a $h(x) < y < v(x)$, donc $\dot{x}$ et $\dot{y}$ sont négatifs ; dans la région (III), $y < h(x)$ donc $\dot{x} < 0$ et $\dot{y} > 0$.



- La région (II) est un piège : dans cette zone, x et y décroissent, donc la solution ne peut s'en échapper ni par en haut (champ vertical sur la frontière), ni par en bas (champ horizontal sur la frontière). Si une

solution a un point dans (II) à un instant t_0, elle y reste donc pour tout $t \geqslant t_0$ et par suite tend vers l'origine.

► Toute solution issue d'un point de (I) ou d'un point de (III) pénètre nécessairement dans (II) et donc tend vers l'origine.

Conclusion : Si $\frac{b_1 c_1}{a_2} \frac{b_2 c_2}{a_1} < 1$, le nombre de sujets infectés tend vers zéro quand t tend vers l'infini : l'épidémie s'arrête.

Le rapport $b_2 c_2 / a_1$ est le nombre moyen de femmes contactées par un homme infecté et en cours de traitement ; on a une interprétation analogue du rapport $b_1 c_1 / a_2$. Ces nombres sont des « taux de contact ». On est amené, en pratique, à partager une population donnée en différentes classes ayant chacune leurs taux de contacts.

6. Étude du moteur électrique

Un moteur électrique se compose essentiellement de deux parties : d'une part un électro-aimant fixe, appelé stator ; d'autre part, une partie mobile, le rotor, situé dans le champ du stator. Le rotor est constitué d'un cylindre porté par l'arbre du moteur et solidaire de bobines alimentées en courant par le collecteur (des bandes latérales isolées les unes des autres sur la paroi du cylindre) au moyen de deux contacts latéraux (les balais).

Les champs magnétiques dans le stator et le rotor tendent à s'aligner : par une habile disposition géométrique des bobinages, on exploite cette propriété de manière à produire un couple moteur sur l'axe du rotor.

Notons θ l'angle entre les champs magnétiques du stator et du rotor. Le couple moteur est une fonction périodique $f(\theta)$ et l'induction électromagnétique crée aussi sur l'axe un couple résistant de moment proportionnel à la vitesse angulaire, donc de la forme $k\dot{\theta}$. En supposant que le couple de charge M est constant, l'équation du mouvement s'écrit

$$I\ddot{\theta} = M - k\dot{\theta} - f(\theta)$$

où I est le moment d'inertie du rotor par rapport à son axe et k une constante positive. En première approximation, la fonction f est sinusoïdale, de la forme $m \sin\theta$, avec $m > 0$. Finalement, l'équation est du type

(1) $$\ddot{\theta} = -a\dot{\theta} - \sin\theta + b$$

où les constantes a et b sont positives. Nous supposons $a > 0$ (dans le cas limite $a = 0$, (1) est une équation de Newton et l'étude se fait comme au chapitre 15).

En posant $v = \dot{\theta}$, on obtient le système différentiel autonome

(2) $$\begin{cases} \dot{\theta} = v \\ \dot{v} = -\sin\theta - av + b \end{cases}$$

Si l'on considère v comme fonction de θ, alors $\dfrac{dv}{d\theta} = \dfrac{\dot{v}}{\dot{\theta}} = \dfrac{\dot{v}}{v}$ et le système (2) s'écrit

$$(1') \qquad v\frac{dv}{d\theta} = -av - \sin\theta + b$$

Le graphe d'une solution de (1') définit une trajectoire de (2) dans le plan des phases (θ,v). L'isocline pour la pente 0, lieu des points (θ,v) où $\dfrac{dv}{d\theta}=0$, est la courbe d'équation $v = \dfrac{b-\sin\theta}{a}$. À un changement d'unité près, celle-ci s'obtient en translatant de b le graphe de $\theta \mapsto -\sin\theta$ le long de l'axe v (figures 1 et 2).

- Si $b > 1$, l'isocline reste au dessus de l'axe des abscisses. On a $\dfrac{dv}{d\theta} > 0$ entre cet axe et l'isocline, et $\dfrac{dv}{d\theta} < 0$ ailleurs ;
- Si $0 < b < 1$, l'isocline coupe l'axe des abscisses : il y a un nombre θ_0 tel que $0<\theta_0<\pi/2$ et $\sin\theta_0=b$. Entre 0 et 2π, les abscisses d'intersection sont θ_0 et $\theta_1=\pi-\theta_0$. On a encore $\dfrac{dv}{d\theta} > 0$ entre l'axe des abscisses et l'isocline et l'on a $\dfrac{dv}{d\theta} < 0$ ailleurs.

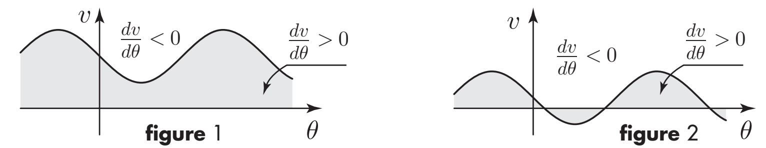


figure 1 **figure** 2

Puisque $\dot{\theta} = v$, le sens de parcours sur une trajectoire est celui pour lequel θ est croissant dans le demi-plan $v > 0$, décroissant dans le demi-plan $v < 0$.
Le second membre de $(1')$ est périodique de période 2π en θ. Il en résulte que si $\theta \mapsto v(\theta)$ est une solution, alors $\theta \mapsto v(\theta + 2\pi)$ est aussi solution.

Les équilibres

Il sont définis par $v = 0 = -av - \sin\theta + b$: ce sont donc les points où l'isocline précédente coupe l'axe des abscisses. Si $b > 1$, il n'y a pas d'équilibre. Si $0 < b < 1$, il y a deux équilibres dans la bande $0 < \theta < 2\pi$: les points $A_0 = (\theta_0, 0)$ et $A_1(\theta_1, 0)$, situés symétriquement par rapport à l'abscisse $\pi/2$.
Si une trajectoire coupe l'isocline ailleurs qu'en un point d'équilibre, elle le fait avec une tangente horizontale ; si une trajectoire coupe l'axe des abscisses, la tangente en ce point est verticale.

Étude dans le cas $b > 1$

Il n'y a pas d'équilibre, mais nous allons montrer que l'équation différentielle $(1')$ possède une solution $\theta \mapsto v(\theta)$ périodique de période 2π. Cela correspond à un mouvement périodique puisqu'à chaque tour de moteur, la vitesse de rotation revient à sa valeur initiale (mais la rotation n'est pas uniforme).

Rappelons que les solutions $v(\theta)$ de $(1')$ sont décroissantes dans leur partie située au dessus de l'isocline $v = \dfrac{b - \sin\theta}{a}$ et qu'elles sont croissantes dans leur partie située entre l'isocline et l'axe des abscisses (figure 1).

Considérons des solutions $v(\theta)$ telles que $v(\pi/2) = v_0 > 0$. Nous allons voir que si v_0 est assez grand, alors on a $v(\pi/2 + 2\pi) < v_0$ et que si v_0 est assez petit, on a au contraire $v(\pi/2 + 2\pi) > v_0$.

► Supposons $v_0 > \dfrac{b+1}{a}$. Si la solution reste constamment au dessus de l'isocline pour $\pi/2 \leqslant \theta \leqslant \pi/2 + 2\pi = 5\pi/2$, elle décroît et l'on a $v(5\pi/2) < v_0$. Sinon, la solution passe sous l'isocline, devient croissante, franchit à nouveau l'isocline, puis décroît jusqu'à l'abscisse $5\pi/2$ (figure 3) ; dans ce cas, $v(\theta)$ n'a pu dépasser le maximum de l'isocline qui a pour valeur $(b+1)/a$: on a donc encore $v(5\pi/2) < v_0$.

► Supposons $0 < v_0 < \dfrac{b-1}{a}$. La solution commence par croître puis franchit l'isocline ; dès lors, la solution reste au dessus de l'isocline jusqu'à $\theta = 5\pi/2$, car le champ de direction le long de l'isocline est horizontal (figure 3). Par suite, $v(5\pi/2)$ est supérieur au minimum $(b-1)/a$ de l'isocline, donc $v(5\pi/2) > v_0$.

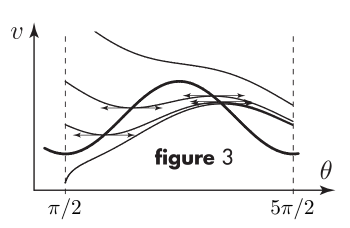
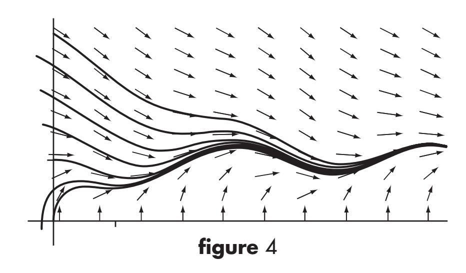


figure 3

Puisque les solutions de $(1')$ dépendent continûment de leur condition initiale, il doit exister une valeur de v_0 pour laquelle $v(5\pi/2) = v_0$.

En effet, considérons la fonction qui à une valeur $v_0 > 0$ associe la différence $\varphi(v_0) = v(5\pi/2) - v_0$, où v est la solution telle que $v(\pi/2) = v_0$. La fonction φ, différence de deux fonctions continues, est continue sur $]0, +\infty[$. On vient de montrer que $\varphi(v_0)$ est négatif si v_0 est assez grand et positif si v_0 est assez petit. D'après le théorème des valeurs intermédiaires, il existe donc une valeur v_0^* telle que $\varphi(v_0^*) = 0$. Pour la solution $\theta \mapsto v^*(\theta)$ telle que $v^*(\pi/2) = v_0^*$, on a ainsi $v^*(\pi/2) = v^*(5\pi/2)$. Le second membre de $(1')$ étant périodique de période 2π en θ, la fonction $w(\theta) = v^*(\theta + 2\pi)$ est solution et comme on a $w(\pi/2) = v^*(\pi/2)$, on en déduit $w(\theta) = v^*(\theta)$ quel que soit θ : autrement dit, la solution v^* est périodique de période 2π.

Dans le mouvement correspondant, le rotor revient dans la même position et à la même vitesse après un tour, donc la rotation se poursuit à l'identique. Cette solution est indiquée sur la figure 3. La figure 4 montre plusieurs trajectoires : on voit qu'elles se rapprochent de la solution périodique quand θ devient grand.


figure 4

Le cylindre des phases. Il est commode de visualiser le mouvement dans le *cylindre des phases* construit de la manière suivante :

- enroulons sur lui-même le plan des phases (θ, v) pour en faire un rouleau cylindrique (comme avec une feuille de papier), de manière que l'axe des θ se transforme en un cercle de périmètre 2π et l'axe des v en une génératrice du cylindre.

Cette opération identifie des points (θ, v) et $(\theta + 2\pi, v)$ correspondant à la même vitesse et à des abscisses qui diffèrent de 2π : dans le cylindre des phases, l'abscisse parcourt un cercle de longueur 2π. Les trajectoires de (2) se transforment en des courbes tracées sur le cylindre, ce qui rend bien compte du mouvement puisque, dans une rotation, θ se compte modulo 2π. La trajectoire correspondant à la solution périodique v^* devient une courbe fermée γ faisant le tour du cylindre : c'est un cycle de fonctionnement régulier du moteur (figure 5).
Sur le cylindre, les trajectoires tendent vers ce cycle limite : cela signifie que le moteur tend vers ce régime de rotation stable.

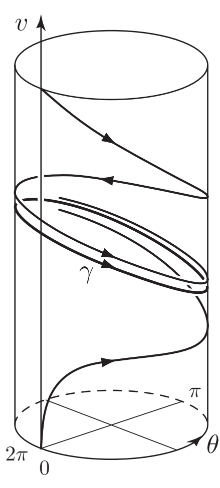


figure 5

Étude des équilibres dans le cas $0 < b < 1$

Pour le système (2), la matrice jacobienne en un point (θ, v) est $J = \begin{bmatrix} 0 & 1 \\ \cos\theta & -a \end{bmatrix}$, de polynôme caractéristique $P = z^2 + az + \cos\theta$. Au point d'équilibre $A_0 = (\theta_0, 0)$, on a $\cos\theta_0 = \sqrt{1-b^2}$; au point $A_1 = (\pi - \theta_0, 0)$, on a $\cos\theta_1 = \cos(\pi - \theta_0) = -\sqrt{1-b^2}$.

Étude au point d'équilibre A_1. Au point A_1, les valeurs propres sont réelles de signes contraires, car leur produit est $\cos\theta_1 < 0$: le point A_1 est un équilibre hyperbolique (équilibre de droite sur les figures 6 et 7).

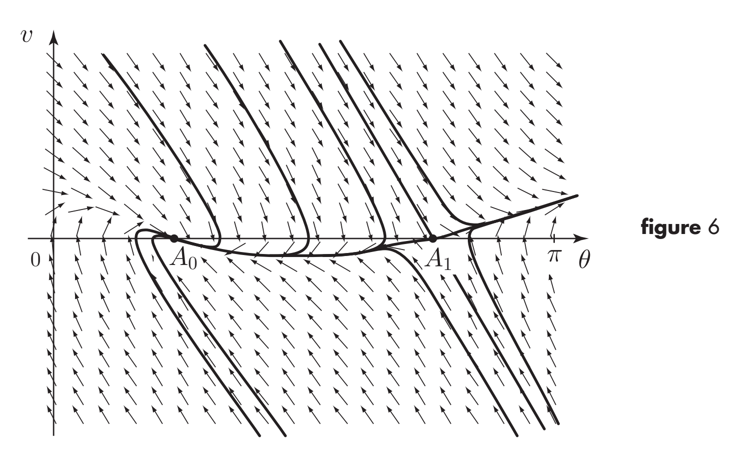


figure 6

Étude au point d'équilibre A_0

- Si $a^2 > 4\sqrt{1-b^2}$, le discriminant de P est positif, donc les valeurs propres sont réelles, de produit $\cos\theta_0 > 0$; elles sont de même signe, celui de leur somme $-a < 0$: les deux valeurs propres sont négatives et par conséquent, les trajectoires tendent vers l'équilibre A_0 (figure 6).
- Si $a^2 < 4\sqrt{1-b^2}$, les racines de P sont complexes conjuguées de partie réelle $-a/2 < 0$: autour de l'équilibre A_0, les trajectoires sont des spirales tendant vers A_0 (figure 7).

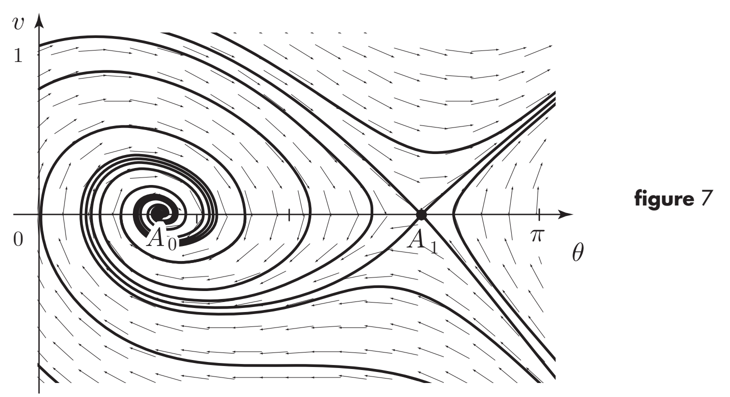


figure 7

Sur les trajectoires qui tendent vers l'équilibre A_0, le moteur « ne fonctionne pas ». Dans certains cas, il n'existe même aucune trajectoire fermée dans le cylindre des phases.
On peut montrer que si a n'est pas trop grand, il existe encore une solution périodique en θ (figure 8). Dans le cylindre des phases, on obtient à nouveau un cycle limite représentant pour le moteur un régime stable.

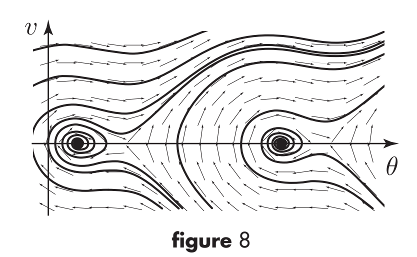


figure 8

7. Une méthode de résolution numérique

Considérons un système différentiel général $X' = F(t, X)$ et la solution $X(t)$ de condition initiale $X(t_0) = X_0$. On cherche à calculer une approximation numérique du vecteur $X(T)$, valeur à l'instant T de la solution.

La méthode d'Euler. Subdivisons l'intervalle d'extrémités t_0 et T en n segments de longueur $|h|$, où $h = (T - t_0)/n$: on introduit donc les points

$$t_0 \ , \ t_1 = t_0 + h \ , \ t_2 = t_0 + 2h \ , \ \dots \ , \ t_n = t_0 + nh = T$$

En t_0, on a $X(t_0) = X_0$ et $X'(t_0) = F(t_0, X_0)$, donc il vient

$$X(t_1) = X(t_0 + h) = X_0 + hF(t_0, X_0) + r(h)$$

où $r(h)$ désigne une quantité vectorielle dont la norme est négligeable devant h quand h tend vers 0.

En posant $Y_1 = X_0 + hF(t_0, X_0)$ on obtient donc une approximation de $X(t_1)$. En t_1, la dérivée $X'(t_1) = F\big(t_1, X(t_1)\big)$ est en général différente de $F\big(t_1, Y_1\big)$, mais pour $|h|$ petit, elle est peu différente : puisque $X(t_2) = X(t_1+h) = X(t_1) + hF\big(t_1, X(t_1)\big) + r_1(h)$ avec $\|r_1(h)\| \underset{h\to 0}{\ll} h$, on prend comme valeur approchée de $X(t_2)$ la quantité $Y_2 = Y_1 + hF(t_1, Y_1)$. Continuant ainsi, on pose $Y_0 = X_0$,

$$Y_1 = Y_0 + hF(t_0, Y_0)\ ,\ \ldots\ ,\ Y_{k+1} = Y_k + hF(t_k, Y_k)\ ,\ \ldots\ ,\ Y_n = Y_{n-1} + hF(t_{n-1}, Y_{n-1})$$

et Y_n est une valeur approchée de $X(T)$. Pour $|h|$ assez petit, l'accumulation des erreurs conduit à une erreur finale $\|X(T) - Y_n\|$ de l'ordre de $|h|$.

Voici une autre méthode numérique, inspirée de la méthode d'Euler, mais un peu plus précise et dont l'implémentation reste simple.

La méthode de la moyenne pente

Expliquons-la dans le cas d'une équation numérique $x' = F(t, x)$: la solution $x(t)$ est donc une fonction à valeurs réelles. En t_0, la pente de la solution est $p = F(t_0, x_0)$ et en t_1, la pente vaut à peu près $q = F(t_1, y_1)$, avec $y_1 = x_0 + hp$. Formons la moyenne $m = \dfrac{p+q}{2}$ des pentes et utilisons m pour estimer une valeur z_1 de $x(t_1)$, en posant

$$z_1 = x_0 + hm$$

On a

$$q = F(t_0 + h, x_0 + hp) = F(t_0, x_0) + h\frac{\partial F}{\partial t}(t_0, x_0) + hp\frac{\partial F}{\partial x}(t_0, x_0) + \rho(h^2)$$

en notant $\rho(h^2)$ une quantité de l'ordre de h^2 quand $|h|$ est assez petit. Il vient

(1)
$$\begin{aligned} z_1 &= x_0 + h\frac{1}{2}\left[p + p + h\frac{\partial F}{\partial t}(t_0, x_0) + hp\frac{\partial F}{\partial x}(t_0, x_0)\right] + \rho(h^3) \\ &= x_0 + hp + \frac{h^2}{2}\left[\frac{\partial F}{\partial t} + p\frac{\partial F}{\partial x}\right](t_0, x_0) + \rho(h^3) \end{aligned}$$

On a aussi le développement limité

$$x(t_1) = x(t_0 + h) = x_0 + hp + \frac{h^2}{2}\frac{\partial^2 x}{\partial t^2}(t_0) + \rho(h^3)$$

Puisque

$$\frac{\partial^2 x}{\partial t^2} = \frac{\partial}{\partial t}\left[F\big(t, x(t)\big)\right] = \frac{\partial F}{\partial t}\big(t, x(t)\big) + \frac{\partial F}{\partial x}\big(t, x(t)\big)x'(t)$$

on obtient $\dfrac{\partial^2 x}{\partial t^2}(t_0) = \left[\dfrac{\partial F}{\partial t} + p\dfrac{\partial F}{\partial x}\right](t_0, x_0)$ et

(2)
$$x(t_1) = x_0 + hp + \frac{h^2}{2}\left[\frac{\partial F}{\partial t} + p\frac{\partial F}{\partial x}\right](t_0, x_0) + \rho(h^3)$$

En comparant (1) et (2), on voit que l'écart $|x(t_1) - z_1|$ est de l'ordre de $|h|^3$.

En continuant ainsi de proche en proche, définissons les approximations $z_0, z_1, \ldots, z_n$ en posant $z_0 = x_0$ et

$$z_{k+1} = z_k + \frac{h}{2}\left[F(t_k, z_k) + F\big(t_{k+1}, z_k + hF(t_k, z_k)\big)\right] \text{ pour } k = 0, 1, \ldots, n-1.$$

Par suite de l'accumulation des erreurs, l'écart final $|x(T) - z_n|$ est seulement, pour $|h|$ assez petit, de l'ordre de h^2.

Algorithme de la moyenne pente

L'algorithme suivant calcule une approximation de $X(T)$, où X est la solution du système différentiel $X' = F(t, X)$ ayant pour condition initiale $X(t_0) = X_0$.
initialisations :

$t \leftarrow t_0$; $Z \leftarrow X_0$; n : un entier positif ; $h \leftarrow (T - t_0)/n$

itération : pour k de 1 à n, faire

$P \leftarrow F(t, Z)$; $s \leftarrow t + h$; $Y \leftarrow Z + hF(t, Z)$; $Q = F(s, Y)$
$Z \leftarrow Z + (h/2)(P + Q)$
$t \leftarrow s$

fin : Z est une approximation de $X(T)$: il y a une constante a telle que $\|X(T) - Z\| \leqslant \frac{a}{n^2}$ pour tout n assez grand.

Il existe bien d'autres méthodes numériques plus précises pour estimer la valeur numérique d'une solution d'équation différentielle, mais les calculs sont plus coûteux.

Application au tracé d'une courbe d'équation $f(x, y) = 0$

Soit f une fonction de deux variables réelles, à dérivées partielles continues, et soit (C) la courbe d'équation $f(x,y)=0$. Supposons qu'on connaisse un point $A=(x_0,y_0)$ sur (C). D'après le théorème des fonctions implicites, on peut paramétrer la courbe (C) au voisinage de A, en prenant par exemple x comme paramètre si la dérivée partielle $\frac{\partial f}{\partial y}$ n'est pas nulle en A. Il y a ainsi une fonction $y(x)$ telle que $f\big(x, y(x)\big) = 0$ pour tout x. En dérivant, on obtient

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = 0 \text{ en tout point de } (C).$$

Ainsi $y(x)$ est la solution de l'équation différentielle (où la variable est x)

$$y' = -\frac{\dfrac{\partial f}{\partial x}(x, y)}{\dfrac{\partial f}{\partial y}(x, y)} \text{ , avec condition initiale } y(x_0) = y_0.$$

En utilisant une méthode numérique, on peut calculer une valeur approchée de $y(x)$ et en traçant le graphe de cette fonction, on obtient l'allure d'un morceau de la courbe (C). S'il y a sur la courbe des points où $\frac{\partial f}{\partial y} = 0$, l'équation différentielle

cesse d'être définie en ces points : le graphe de la fonction $y(x)$ peut donc ne pas couvrir en totalité la courbe (C).

Exercices

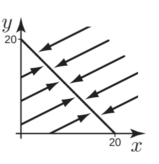
@ **1. Évolution de taux d'intérêts.** Reprenons le système différentiel $\begin{cases} \dot{c} = a(r-c-\ell) \\ \dot{\ell} = b(r-c-\ell) \end{cases}$ de l'exemple **2** page 479 qui modélise l'évolution des taux d'intérêts c et ℓ à court et à long terme (les constantes a, b, r sont positives). Le domaine est $c > 0$ et $\ell > 0$.

a) Montrer que tout point du segment de droite d'équation $c + \ell = r, c > 0, r > 0$ est un point d'équilibre et que si $\big(c(t), \ell(t)\big)$ est solution, alors $bc(t) - a\ell(t)$ est constant. En déduire que les trajectoires sont rectilignes.

b) Montrer que la fonction constante $(c,\ell) = \left(\frac{ar}{a+b}, \frac{br}{a+b}\right)$ est une solution du système.

c) Calculer les valeurs propres et les vecteurs propres de la matrice du système. En déduire la solution générale.

d) Soit $S(t) = \big(c(t), \ell(t)\big)$ la solution ayant pour condition initiale $c(0) = c_0$, $\ell(0) = \ell_0$, où c_0 et ℓ_0 sont des nombres positifs. Calculer $S(t)$ et montrer que $S(t)$ tend vers le point d'équilibre $\frac{1}{a+b}(bc_0 - a\ell_0 + ar, -bc_0 + a\ell_0 + br)$ quand t tend vers $+\infty$. Montrer que chaque point d'équilibre est stable, mais pas asymptotiquement stable.



e) Justifier le dessin ci-contre des trajectoires, où l'on a choisi $a=2/3$, $b=1/3$ et $r=20$.

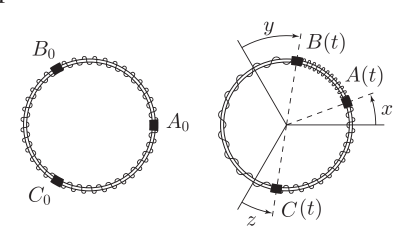
2. Un système de ressorts. Dans l'exemple **1** page 478, prenons des masses unité et des ressorts de raideur $k_1 = 1$, $k_2 = 2$ et $k_3 = 4$. Le système différentiel s'écrit $X' = AX$, où $A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -3 & 2 & 0 & 0 \\ 2 & -6 & 0 & 0 \end{bmatrix}$.

a) Montrer que les valeurs propres de A sont $\pm\sqrt{2}$ et $\pm\sqrt{7}$.

b) Calculer une base de solutions et montrer que le mouvement des masses est donné par $\begin{cases} x_1(t) = a\cos(\sqrt{2}\,t) + b\sin(\sqrt{2}\,t) + c\cos(\sqrt{7}t) + d\sin(\sqrt{7}\,t) \\ x_2(t) = \frac{a}{2}\cos(\sqrt{2}\,t) + \frac{b}{2}\sin(\sqrt{2}\,t) - 2c\cos(\sqrt{7}\,t) - 2d\sin(\sqrt{7}\,t) \end{cases}$ où a, b, c, d sont des constantes.

c) À l'instant $t = 0$, on maintient la masse m_2 à l'équilibre et l'on écarte m_1 de la quantité algébrique $a \neq 0$. Montrer que le mouvement est donné par $\frac{x_1}{a} = \frac{4}{5}\sin\sqrt{2}t + \frac{1}{5}\cos\sqrt{7}t$ et $\frac{x_2}{a} = \frac{2}{5}\sin\sqrt{2}t - \frac{2}{5}\cos\sqrt{7}t$. Ces mouvements sont-ils périodiques ? (procéder comme dans l'exercice **5** page 294).

3. Ressorts en anneau. On enfile trois ressorts identiques sur un anneau circulaire rigide R placé horizontalement, chaque ressort ayant pour longueur un tiers de circonférence. Appelons A_0, B_0, C_0 leurs extrémités. Quand on abandonne le dispositif après avoir comprimé un ou plusieurs ressorts, les extrémités des ressorts ont des mouvements $A(t), B(t), C(t)$ qu'on repère par les arcs orientés $x = \overparen{A_0A}$, $y = \overparen{B_0B}$ et $z = \overparen{C_0C}$.



a) Notons ℓ_1, ℓ_2, ℓ_3 la longueur des ressorts AB, BC, CA. Montrer qu'à tout instant, on a $x+\ell_1 - y = y+\ell_2-z = z+\ell_3-x$.

b) Montrer que $\ddot{x} = -k(L-\ell_1) + k(L-\ell_3) = -k(2x-y-z)$, où k est la raideur des ressorts et L le tiers de la circonférence. Écrire les égalités analogues pour $\ddot{y}$ et $\ddot{z}$.

c) Supposons $k=1$ et posons $u=\dot{x}$, $v=\dot{y}$, $w=\dot{z}$. Écrire le système différentiel linéaire $X' = AX$ aux six inconnues x, y, z, u, v, w. Montrer que la matrice A possède trois valeurs propres doubles 0, $\boldsymbol{i}\sqrt{3}$ et $-\boldsymbol{i}\sqrt{3}$.

d) Supposons que les positions initiales sont données par $x(0) = 0, y(0) = b, z(0) = c$ et qu'on abandonne les ressorts sans vitesse initiale. Montrer que les solutions sont $x(t) = p\big[1-\cos(\sqrt{3}\,t)\big]$, $y(t) = p + \dfrac{2b-c}{3}\cos(\sqrt{3}\,t)$, $z(t) = p + \dfrac{2c-b}{3}\cos(\sqrt{3}\,t)$, où $p = \dfrac{b+c}{3}$. Les mouvements des points A, B, C sont-ils périodiques ?

@ **4. Second membre exponentiel.** Soit un système différentiel linéaire $X' = AX + B(t)$, où A est une matrice constante de taille n.

a) Supposons que $B(t)=e^{\alpha t}V$, où V est un vecteur constant et α un nombre qui n'est pas valeur propre de A. Montrer qu'il y a un unique vecteur Y solution de l'équation linéaire $(\alpha I_n - A)Y = V$ et que $e^{\alpha t}Y$ est une solution du système différentiel.

b) La diffusion d'une certaine substance s entre trois compartiments tissulaires (1), (2) et (3) est modélisée, en régime libre, par le système $X'=AX$, où $A=\begin{bmatrix} -2 & 1/2 & 1 \\ 1 & -2 & 1 \\ 1 & 3/2 & -2 \end{bmatrix}$.

Initialement, les compartiments ne contiennent pas s. À l'instant $t = 0$, on injecte s dans le compartiment (1) selon la loi de diffusion $pe^{-t} + qe^{-2t}$, où p et q sont des quantités positives. Montrer qu'à l'instant t, la quantité de substance s dans le compartiment (2) est $y(t) = \dfrac{1}{6}\big[2p+q - 3pe^{-t} - 3qe^{-2t} + (p+2q)e^{-3t}\big]$.
Quelle est l'allure de la courbe $y(t)$ pour $t \geqslant 0$?

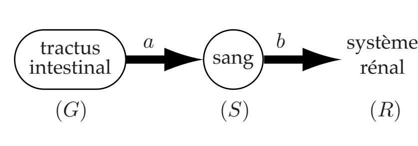
@ **5. L'équilibre proies-prédateurs.** Considérons le système proies-prédateurs étudié page 501.

a) Calculer le linéarisé du système au point d'équilibre $(0,0)$ et justifier la forme des trajectoires autour de ce point.

b) Calculer le linéarisé (L) : $X'=AX$ du système au point d'équilibre $E=(c/d, a/b)$ et montrer que le système n'est pas linéarisable en ce point.

c) Quelle approximation peut-on prendre pour une solution de condition initiale voisine de E ? Montrer que la période des petites oscillations autour de E est à peu près $2\pi/\sqrt{ac}$ (voir le paragraphe **4**).

@ **6. Dépendance des paramètres.** Considérons le système compartimental formé du tractus intestinal (G) et du sang (S) : les transferts de substances de (G) vers (S) et de (S) vers le système rénal (R) sont régis par la loi d'action de masse : $\begin{cases} \dot{G} = -aG \\ \dot{S} = aG - bS \end{cases}$, où a et b sont des paramètres positifs tels que $b > a$.



a) On suppose qu'à l'instant $t = 0$, on a $G(0) = g_0 > 0$ et $S(0) = 0$. Calculer la solution $\big(G(t), S(t)\big)$. Montrer que $G(t)$ et $S(t)$ tendent vers 0 quand t tend vers $+\infty$ et que l'équilibre $(0, 0)$ est asymptotiquement stable.

b) Les fonctions de sensibilité au paramètre a sont $x(t) = \dfrac{\partial G}{\partial a}$ et $y(t) = \dfrac{\partial S}{\partial a}$. Montrer que ces fonctions sont solutions du système $\begin{cases} \dot{x} = -ax - G(t) \\ \dot{y} = ax - by + G(t) \end{cases}$. Montrer que $x(t) = -g_0 t e^{-at}$ et $y(t) = \dfrac{g_0}{(b-a)^2}\Big[\big[b - a(b-a)t\big]e^{-at} - be^{-bt}\Big]$.

c) On veut estimer le paramètre de diffusion a en mesurant la quantité S par des analyses sanguines. On aura une meilleure précision si l'on effectue ces mesures à un instant où a dépend le moins de S, c'est-à-dire lorsque la valeur absolue $\left|\dfrac{\partial a}{\partial S}\right|$ est minimum. Supposons que la valeur de a est proche de 1 et que celle de b est proche de 1,5. Pour ces valeurs, faire dessiner par ordinateur le graphe de $y(t)$ entre 0 et 6. Constater qu'il vaut mieux faire les mesures vers l'instant $t = 3$.

@ **7. Un exemple de compétition animale.** Le système suivant modélise une compétition brutale entre deux populations animales d'effectifs x et y (l'unité est la dizaine d'individus) : $\begin{cases} \dot{x} = 2x - xy \\ \dot{y} = a(y - y^2 - xy) \end{cases}$, où $a > 0$. Le terme $-y^2$ reflète une mortalité par surpopulation et $-xy$ traduit les contacts mortels entre individus d'espèces différentes.

a) Trouver les deux points d'équilibre à coordonnées positives ou nulles. Montrer que le demi-axe des $y \geqslant 0$ est réunion de quatre trajectoires. Montrer que le demi-axe $x > 0$ est une trajectoire : quelle sont les solutions $\big(x(t), 0\big)$ correspondantes ?

b) Montrer que les droites d'équation $y = 2$ et $x + y = 1$ divisent le domaine $D = \{(x,y) \mid x > 0 \text{ et } y > 0\}$ en trois régions dans lesquelles $\dot{x}$ et $\dot{y}$ ont des signes constants. On note (A) la région $y > 2$, (C) la région $x + y < 1$ et (B) la troisième région. Dessiner la direction du champ en tout point d'une frontière de ces régions.

c) Montrer que toute solution qui part d'un point de (A) ou d'un point de (C) rentre dans la région (B). Montrer que toute solution qui est dans (B) à un certain instant reste ultérieurement dans (B).

d) Montrer que pour toute solution, on a $\lim_{t\to+\infty} x(t) = +\infty$. En déduire que $y(t)$ a une limite finie comprise entre 0 et 2 quand t tend vers $+\infty$.

e) Montrer qu'il existe un instant T tel que l'on ait $\dot{y}(t) \leqslant -y(t)$ quel que soit $t \geqslant T$. En déduire que $y(t)$ tend vers 0 quand t tend vers $+\infty$.

f) Faire un dessin des trajectoires.

g) On suppose $a = 1$. Soit $\big(x(t), y(t)\big)$ la solution telle que $x(0) = 2$ et $y(0) = 5$. On considère que la population y est éteinte lorsque $y/x \leqslant 0{,}05$. En faisant calculer numériquement les fonctions $x(t)$ et $y(t)$ par un ordinateur, montrer que le temps d'extinction de y est environ $1{,}19$.

8. L'équation de Van der Pol. Il s'agit de l'équation différentielle (P) : $y'' = -y - y' + y'^3$. Écrire le système différentiel du premier ordre associé. En utilisant un ordinateur, vérifier que les trajectoires ont l'allure ci-contre. Expliquer comment, par ce dessin, on peut se convaincre qu'il y a une solution périodique. Cette équation a été utilisée pour les premières modélisations du fonctionnement cardiaque.

@ **9. Utilisation de la formule de Stokes.** Soit le système différentiel $\begin{cases} x' &= f(x,y) \\ y' &= g(x,y) \end{cases}$, défini dans un domaine D (les fonctions f et g ont des dérivées partielles continues). Si R est une région de D limitée par une courbe fermée C, la formule de Stokes (page 430) affirme que l'on a $\int_C f(x,y)\,dy - g(x,y)\,dx = \iint_R \left[\frac{\partial f}{\partial x}(x,y) + \frac{\partial g}{\partial y}(x,y)\right] dx\,dy$.

a) Supposons que $\big(x(t), y(t)\big)$ est une solution périodique du système différentiel et que C est sa trajectoire. Montrer que, dans la formule de Stokes, l'intégrale curviligne est nulle.

b) Supposons que le champ de vecteurs $\vec{E} = \big[f(x,y), g(x,y)\big]$ a une divergence de signe constant dans D. Montrer que le système différentiel n'a aucune solution périodique.

c) Considérons le système différentiel $\begin{cases} \dot{\theta} &= v \\ \dot{v} &= -\sin\theta - av + b \end{cases}$ du moteur électrique, où les constantes a et b sont strictement positives (page 514). Montrer que la divergence du champ est $-a$. En déduire que dans le cylindre des phases, toute trajectoire fermée doit faire le tour du cylindre.

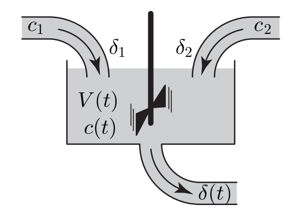
@ **10. Étude d'un équilibre.** Soient $a(x)$ et $c(x)$ des fonctions à dérivées continues, définies sur un intervalle autour de 0. Considérons le système différentiel $\begin{cases} x' &= xa(x) + by \\ y' &= xc(x) + dy \end{cases}$, où b et d sont des constantes.

a) Montrer que l'origine $O = (0,0)$ est un point d'équilibre et que le linéarisé du système en ce point a pour matrice $\begin{bmatrix} a(0) & b \\ c(0) & d \end{bmatrix}$.

b) En déduire que si l'on a $a(0)d - bc(0) > 0$ et $a(0) + d < 0$, alors l'origine est asymptotiquement stable.

c) Supposons que pour tout x assez voisin de 0 et différent de 0, on a $a(x)d - bc(x) > 0$ et $a(x) + d < 0$. Supposons de plus $b \neq 0$. Posons $V(x,y) = \frac{1}{2}(dx - by)^2 + \int_0^x \big[a(s)d - c(s)b\big]\, s\, ds$. Montrer que V est une fonction de Liapounov pour l'équilibre O et que cet équilibre est asymptotiquement stable.

11. Contrôle d'une cuve à mélanger. Considérons une cuve mélangeuse contenant un solvant et alimentée par deux solutions de concentrations c_1 et c_2, avec des débits réglables δ_1 et δ_2. Soit $V(t)$ le volume de liquide dans la cuve à l'instant t, $\delta(t)$ le débit de sortie et $c(t)$ la concentration du mélange dans la cuve.



Les bilans en volume de solution et en masse de produits dissous s'expriment par

$$\frac{dV}{dt} = \delta_1 + \delta_2 - \delta \quad \text{et} \quad \frac{d(cV)}{dt} = c_1\delta_1 + c_2\delta_2 - c\delta\,.$$

De plus, si S est la section de la cuve, alors d'après la loi d'écoulement de Bernouilli, on a $\delta = a\sqrt{V}$ (où la constante positive a est de la forme $b/\sqrt{S}$).

a) Montrer qu'un état d'équilibre est caractérisé par un volume constant V_e dans la cuve ainsi que par des débits $\delta_1 = \delta_{1e}$, $\delta_2 = \delta_{2e}$, $\delta = \delta_e$ et une concentration en sortie $c = c_e$ vérifiant les égalités $\delta_e = \delta_{1e} + \delta_{2e}$ et $c_e\delta_e = c_1\delta_{1e} + c_2\delta_{2e}$.

b) Pour linéariser le système autour de l'équilibre, introduisons les différences $x = V - V_e$ et $y = c - c_e$ et les commandes $u = \delta_1 - \delta_{1e}$, $v = \delta_2 - \delta_{2e}$.

(i) Montrer que pour x petit, on a au premier ordre l'approximation

$$\sqrt{V_e + x} \sim \sqrt{V_e}\left(1 + \frac{x}{2V_e}\right).$$

(ii) En déduire les approximations

$$\dot{x} = \frac{-ax}{2\sqrt{V_e}} + u + v \quad \text{et} \quad \dot{y} = \frac{-c}{V_e}[u + v + \delta_e] + \frac{1}{V_e}[c_1u + c_2v + c_e\delta_e]\,.$$

(iii) Posons $p = \frac{a}{2\sqrt{V_e}}$, $\alpha_1 = \frac{c_1 - c_e}{V_e}$ et $\alpha_2 = \frac{c_2 - c_e}{V_e}$. Montrer que le linéarisé du système au point d'équilibre est (L) : $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -p & 0 \\ 0 & -2p \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$

c) Calculer la matrice de commandabilité de (L). En déduire que le système (L) est commandable par (u,v) si et seulement si les concentrations c_1 et c_2 sont différentes. Justifier l'affirmation : « dans le cas $c_1 \neq c_2$, si deux états (V,c) et (V',c') sont proches de l'équilibre, on peut passer du premier au second en agissant sur les débits δ_1 et δ_2. »

d) Soit la matrice $K = \begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix}$. Calculer la matrice $A - BK$ et son polynôme caractéristique. On suppose $c_1 \neq c_2$ et l'on pose $k = \varepsilon/(2\alpha_2)$, où $\varepsilon = 1$ si $\alpha_2 - \alpha_1 > 0$ et $\varepsilon = -1$ si $\alpha_2 - \alpha_1 < 0$. Montrer que le système $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = (A - BK)\begin{bmatrix} x \\ y \end{bmatrix}$ est un bouclage stabilisant de (L).

@12. Soit $X' = f(t, X)$ un système différentiel où la fonction f a des dérivées partielles continues. On suppose que f est périodique de période T par rapport à la variable t, autrement dit $f(t+T, X) = f(t, X)$ pour tous t et X.
Montrer que si $t \mapsto X(t)$ est une solution, alors la fonction $t \mapsto X(t+T)$ est solution. En déduire que s'il existe t_0 tel que $X(t_0) = X(t_0+T)$, alors la fonction $X(t)$ est périodique de période T (comparer les conditions initiales des solutions $X(t)$ et $X(t+T)$).

Chapitre 17

Séries, séries entières, séries de Fourier

1. Séries numériques

Étant donnée une suite de nombres u_k réels ou complexes, on doit parfois considérer les sommes $s_0 = u_0$, $s_1 = u_0 + u_1$, $s_2 = u_0 + u_1 + u_2$ et de manière générale, la somme d'un nombre quelconque de premiers termes :

$$s_n = u_0 + u_1 + \cdots + u_n = \sum_{k=0}^{n} u_k$$

Définition

Les sommes s_n forment une suite, appelée *série de terme général* u_k et notée $\sum u_k$. Chaque somme s_n s'appelle une *somme partielle* de la série.

Si les u_k ne sont définis que pour $k \geqslant p$, les sommes partielles sont

$$u_p \ , \ u_p + u_{p+1} \ , \ \ldots \ , \ u_p + u_{p+1} + \cdots + u_n = \sum_{k=p}^{n} u_k \quad \text{pour tout } n \geqslant p.$$

On s'intéresse surtout aux cas où les sommes partielles forment une suite convergente.

Définition

Si la suite s_n a une limite s, on dit que *la série* $\sum u_n$ *est convergente*. Dans ce cas, le nombre s s'appelle *la somme* de la série et se note $s = \sum_{k=p}^{+\infty} u_k$, où p est le premier indice de définition des u_k. Les nombres $r_n = s - s_n$ s'appellent les *restes* de la série.

Si des séries $\sum u_k$ et $\sum u'_k$ sont convergentes de sommes s et s', alors la série $\sum (u_k + u'_k)$ a pour somme $s + s'$, et pour tout nombre λ, la série $\sum (\lambda u_k)$ a pour somme λs.
Par définition, une série à termes complexes est convergente si et seulement si la série des parties réelles et la série des parties imaginaires sont toutes deux convergentes.

Proposition. *Si une série* $\sum u_k$ *est convergente, alors* $\lim_{k \to +\infty} u_k = 0$.

Démonstration. En effet, en supposant que la suite des u_k commence à l'indice 0, on a $s_n - s_{n-1} = u_n$ pour tout $n \geqslant 1$. Si les sommes s_n tendent vers s, alors la différence $s_n - s_{n-1}$ tend vers $s - s = 0$. ■

Dans le cas d'une série $\sum u_k$ à termes u_k réels positifs ou nuls, les sommes partielles sont croissantes puisqu'on a $s_{n+1} - s_n = u_n \geqslant 0$ pour tout n. Comme une suite croissante de nombres réels est convergente si et seulement si elle est majorée (page 261), on en déduit le résultat suivant.

Théorème. *Une série $\sum u_k$ à termes $u_k \geqslant 0$ est convergente si et seulement si ses sommes partielles sont majorées.*

Théorème de comparaison. *Soient $\sum u_k$ et $\sum v_k$ des séries à termes réels. Si l'on a $0 \leqslant u_k \leqslant v_k$ pour tout k et si la série $\sum v_k$ converge, alors la série $\sum u_k$ converge.*

Démonstration. Chaque somme partielle de la série $\sum u_k$ est inférieure ou égale à la somme partielle correspondante de la série $\sum v_k$. Si cette dernière série converge, ses sommes partielles sont majorées, donc aussi les sommes partielles de la série $\sum u_k$ qui est donc convergente. ■

1.1 Les exemples fondamentaux

Séries géométriques

Supposons que les u_k forment une suite géométrique : si a est le premier terme et z la raison, on a donc $u_k = az^k$ quel que soit l'entier $k \geqslant 0$. On dit que la série $\sum az^k$ est une *série géométrique*. Les sommes partielles sont (remarque page 46)

$$s_n = az^0 + az^1 + az^2 + \cdots + az^n = a[1 + z + z^2 + \cdots + z^n] = a\frac{1-z^{n+1}}{1-z}$$

Proposition. *Supposons $a \neq 0$. La série géométrique $\sum az^k$ est convergente si et seulement si $|z| < 1$ et dans ce cas, sa somme est $\sum_{k=0}^{+\infty} az^k = \frac{a}{1-z}$.*

Démonstration. Si $|z| < 1$, alors $|z^n| = |z|^n$ tend vers 0, donc aussi z^n : quand n tend vers $+\infty$, les sommes s_n ont alors pour limite $a\frac{1-0}{1-z} = \frac{a}{1-z}$ et la série est convergente. Réciproquement, si la série $\sum az^k$ converge, alors nécessairement son terme général az^k tend vers 0 quand k tend vers $+\infty$, donc aussi les modules $|az^k| = |a||z|^k$; puisqu'on a supposé $a \neq 0$, $|z|^k$ tend vers 0 et cela implique $|z| < 1$. ■

Majoration du reste lorsque $|z| < 1$. En notant s la somme de la série, on a $s_n = s - a\frac{z^{n+1}}{1-z}$, donc le reste est $r_n = s - s_n = a\frac{z^{n+1}}{1-z}$. D'après l'inégalité triangulaire $1 \leqslant |z| + |1-z|$, on a $|1-z| \geqslant 1 - |z|$, d'où la majoration $|r_n| \leqslant \frac{|a|}{1-|z|}|z|^{n+1}$ qui est d'autant plus petite que n est grand.

Séries de Riemann

Définition

Soit α un nombre réel positif. La série $\sum \frac{1}{k^\alpha}$ s'appelle une *série de Riemann*.

Pour étudier la convergence, nous allons estimer les sommes partielles $1+\frac{1}{2^\alpha}+\cdots+\frac{1}{n^\alpha}$ en les comparant à une intégrale de la fonction $x \mapsto \frac{1}{x^\alpha}$.

Proposition. *Soient $p \geqslant 0$ un entier et $f : [p, +\infty[\to \mathbb{R}$ une fonction continue positive et décroissante. Alors la série $\sum f(k)$ est convergente si et seulement si l'intégrale généralisée $\int_p^{+\infty} f(t)\,dt$ existe. Dans ce cas, on a l'encadrement*

$$\int_p^{+\infty} f(t)\,dt \leqslant \sum_{k=p}^{+\infty} f(k) \leqslant f(p) + \int_p^{+\infty} f(t)\,dt.$$

Démonstration. Pour tout entier $k \geqslant p$, posons $u_k = f(k)$. Si t est entre deux entiers k et $k+1$, alors on a $u_{k+1} = f(k+1) \leqslant f(t) \leqslant f(k) = u_k$, car f est décroissante. En intégrant sur $[k, k+1]$, il vient donc $u_{k+1} \leqslant \int_k^{k+1} f(t)\,dt \leqslant u_k$. Ajoutons ces inégalités pour $k = p, p+1, \ldots, n$; on obtient :

$$u_{p+1} + u_{p+2} + \cdots + u_{n+1} \leqslant \int_p^{n+1} f(t)\,dt \leqslant u_p + u_{p+1} + \cdots + u_n$$

Le terme de droite est la somme partielle s_n de la série $\sum u_k$ de premier terme u_p et celui de gauche est $s_{n+1} - u_p$. L'encadrement se récrit ainsi sous la forme

$$(*) \qquad \int_p^{n+1} f(t)\,dt \leqslant s_n \leqslant u_p + \int_p^n f(t)\,dt$$

où l'on a changé $n+1$ en n pour avoir l'inégalité de droite. Supposons que l'intégrale généralisée $I = \int_p^{+\infty} f(t)\,dt$ existe : par définition, I est la limite des intégrales $\int_p^n f(t)\,dt$ quand n tend vers l'infini. Puisque la fonction f est positive, on a $\int_p^n f(t)\,dt \leqslant I$ pour tout n, donc les sommes partielles s_n sont majorées par le nombre $u_p + I$. D'après le théorème précédent, la série $\sum u_k$ converge. Réciproquement, si la série converge, ses sommes partielles sont majorées par la somme s de la série et les intégrales $\int_p^{n+1} f(t)\,dt$ sont, d'après $(*)$, majorées par s. Il s'ensuit que l'on a $\int_p^x f(t)\,dt \leqslant s$ pour tout $x \geqslant p$ et par suite, l'intégrale généralisée existe. En passant à la limite dans $(*)$ quand n tend vers l'infini, on obtient l'encadrement pour la somme s. ■

On sait que l'intégrale généralisée $\int_1^{+\infty} \frac{dt}{t^\alpha}\,dt$ existe si et seulement si $\alpha > 1$ (page 325). D'après la proposition, c'est aussi la condition de convergence de la série de Riemann $\sum \frac{1}{k^\alpha}$.

Proposition. *La série de Riemann $\sum \frac{1}{k^\alpha}$ est convergente si et seulement si $\alpha > 1$.*

Majoration du reste. Si la série satisfait les hypothèses de la proposition, l'encadrement donné permet de majorer le reste $r_n = s - s_n$ pourvu qu'on puisse majorer, en fonction de n, l'intégrale généralisée $\int_{n+1}^{+\infty} f(t)\, dt$.

Exemple. Considérons la série de terme général $u_k = \dfrac{1}{a^2+k^2}$, où $a > 0$. La fonction $f(x) = \dfrac{1}{a^2+x^2}$ est positive et décroissante sur $[0, +\infty[$, donc pour tout entier $p \geqslant 1$, on a la majoration

$$\sum_{k=p}^{+\infty} \frac{1}{a^2+k^2} \leqslant u_p + \int_p^{+\infty} \frac{dt}{a^2+t^2}\,.$$

Puisque $\displaystyle\int \frac{dt}{a^2+t^2} = \frac{1}{a} \operatorname{Arc\,tan} \frac{x}{a}$ (page 298), il vient $\displaystyle\int_p^{+\infty} \frac{dt}{a^2+t^2} = \frac{1}{a}\left[\frac{\pi}{2} - \operatorname{Arc\,tan} \frac{p}{a}\right]$.
Mais d'après la relation $\tan\left(\dfrac{\pi}{2} - \theta\right) = \dfrac{1}{\tan\theta}$, on a $\dfrac{\pi}{2} - \operatorname{Arc\,tan} x = \operatorname{Arc\,tan} \dfrac{1}{x}$ pour tout $x > 0$. Ainsi, $\displaystyle\int_p^{+\infty} \frac{dt}{a^2+t^2} = \frac{1}{a} \operatorname{Arc\,tan} \frac{a}{p}$ et

$$\sum_{k=p}^{+\infty} \frac{1}{a^2+k^2} \leqslant \frac{1}{a^2+p^2} + \frac{1}{a} \operatorname{Arc\,tan} \frac{a}{p}.$$

Cette majoration du reste permet un calcul approché de la somme de la série, car le membre de droite M_p tend vers 0 quand p tend vers l'infini : si l'on trouve un entier $p \geqslant 2$ assez grand pour que $M_p < \varepsilon$, alors la somme partielle $u_1 + u_2 + \cdots + u_{p-1}$ est une valeur approchée par défaut à ε près de la somme $\sum_{k=1}^{+\infty} u_k$.

Séries d'Abel

Il s'agit des séries à termes complexes de la forme $\sum a_k e^{i\,kx}$, où le nombre x et les a_k sont réels. La série converge si et seulement si chacune des séries $\sum a_k \sin kx$ et $\sum a_k \cos kx$ est convergente. On a le résultat suivant, dont nous verrons des exemples dans le cadre des séries de Fourier.

Proposition. *Si la suite des a_k est décroissante, positive et tend vers 0, alors la série $\sum a_k \sin kx$ est convergente pour tout x et la série $\sum a_k \cos kx$ est convergente au moins pour x non multiple entier de 2π.*

Exemples

- **Séries alternées.** En prenant $x = \pi$, on a $\cos kx = \cos k\pi = (-1)^k$ pour k entier.

  Si les a_k sont positifs, décroissants et tendent vers 0, la série $\sum (-1)^k a_k$ est convergente.

  Une telle série est dite *alternée*. Ainsi la série alternée $\sum \dfrac{(-1)^k}{k}$ est convergente, de même que $\sum \dfrac{(-1)^k}{k^\alpha}$ pour tout nombre $\alpha > 0$.
- Si $\alpha > 0$, la série $\sum \dfrac{\sin kx}{k^\alpha}$ est convergente quel que soit x.

Mode de convergence d'une série alternée. Les sommes partielles sont

$$s_n = a_0 - a_1 + a_2 - a_3 + a_4 - a_5 + \cdots + (-1)^n a_n$$

On a

$$s_2 - s_0 = -a_1 + a_2 \leqslant 0 \ , \quad s_4 - s_2 = -a_3 + a_4 \leqslant 0 \ , \quad s_{2n} - s_{2n-2} = -a_{2n-1} + a_{2n} \leqslant 0$$
$$s_3 - s_1 = a_2 - a_3 \geqslant 0 \ , \quad s_5 - s_3 = a_4 - a_5 \geqslant 0 \ , \quad s_{2n+1} - s_{2n-1} = a_{2n} - a_{2n+1} \geqslant 0$$

Les sommes $s_1, s_3, s_5, \ldots$ forment une suite croissante, les sommes $s_0, s_2, s_4, \ldots$ une suite décroissante, et la différence $s_{2n} - s_{2n-1} = a_{2n}$ est positive et tend vers 0 : ainsi la suite des s_{2n-1} et la suite des s_{2n} sont adjacentes, donc elles ont la même limite s (page 261) : cela démontre qu'une série alternée est convergente. De plus, on a l'encadrement $s_{2n-1} \leqslant s \leqslant s_{2p}$ pour tout $n \geqslant 1$ et $p \geqslant 0$.

La somme d'une série alternée est toujours comprise entre deux sommes partielles consécutives quelconques.

Majoration du reste d'une série alternée

En appelant s la somme de la série, on a $0 \leqslant s - s_{2n-1} \leqslant s_{2n} - s_{2n-1} = a_{2n}$ et $0 \leqslant s_{2n} - s \leqslant s_{2n} - s_{2n+1} = a_{2n+1}$.

Dans une série alternée, le reste est, en valeur absolue, inférieur au premier terme négligé.

1.2 Critères de convergence

Proposition et définition. *Si la série $\sum |u_k|$ est convergente, alors la série $\sum u_k$ est convergente aussi et $\left|\sum_{k=0}^{+\infty} u_k\right| \leqslant \sum_{k=0}^{+\infty} |u_k|$. Dans ce cas, on dit que la série $\sum u_k$ est absolument convergente.*

Démonstration. Notons a_k la partie réelle de u_k et b_k la partie imaginaire, de sorte que $u_k = a_k + \boldsymbol{i}\, b_k$. On sait que $|a_k| \leqslant |u_k|$ et $|b_k| \leqslant |u_k|$. Si la série $\sum |u_k|$ est convergente, alors d'après le théorème de comparaison, il en va de même des séries $\sum |a_k|$ et $\sum |b_k|$. Puisqu'on a $0 \leqslant |a_k| - a_k \leqslant 2|a_k|$, la série de terme général $d_k = |a_k| - a_k$ est convergente. On a $\sum a_k = \sum |a_k| - \sum d_k$, donc la série $\sum a_k$ est convergente. De même, la série $\sum b_k$ est convergente et il s'ensuit que la série $\sum (a_k + \boldsymbol{i}\, b_k) = \sum u_k$ est convergente.
Pour toute somme partielle $s_n = \sum_{k=0}^{n} u_k$, on a $|s_n| \leqslant \sum_{k=0}^{n} |u_k|$ car le module d'une somme est inférieur ou égal à la somme des modules ; par suite,

$$\left|\sum_{k=0}^{+\infty} u_k\right| = |\lim_{n \to +\infty} s_n| = \lim_{n \to +\infty} |s_n| \leqslant \sum_{k=0}^{+\infty} |u_k| \, .$$ ■

En appliquant le théorème de comparaison, on en déduit le premier critère suivant.

Critère de comparaison. *Si l'on a $|u_k| \leqslant v_k$ pour tout k assez grand et si la série $\sum v_k$ converge, alors la série $\sum u_k$ est absolument convergente (donc convergente).*

Exemples

► La série alternée $\sum \frac{(-1)^k}{k}$ est convergente, mais elle n'est pas absolument convergente, car $\left|\frac{(-1)^k}{k}\right| = \frac{1}{k}$ est le terme général d'une série de Riemann non convergente. De même, la série $\sum \frac{(-1)^k}{\sqrt{k}}$ est convergente, mais pas absolument convergente.

► Si $\alpha > 1$, la série $\sum \frac{e^{\boldsymbol{i}\,kx}}{k^\alpha}$ est absolument convergente pour tout réel x : en effet, on a $\left|\frac{e^{\boldsymbol{i}\,kx}}{k^\alpha}\right| = \frac{1}{k^\alpha}$ et la série de Riemann $\sum \frac{1}{k^\alpha}$ est convergente.

Proposition. *La somme de deux séries absolument convergentes est absolument convergente.*

On a en effet $|u_k+v_k| \leqslant |u_k|+|v_k|$. Si $\sum |u_k|$ et $\sum |v_k|$ convergent, alors leur somme $\sum \big(|u_k|+|v_k|\big)$ converge et, d'après le critère de comparaison, il en va de même de $\sum |u_k+v_k|$.

Critères de convergence pour les séries à termes réels positifs

Supposons $u_k > 0$ pour tout k.

a) *Si l'on a $u_k \underset{k\to+\infty}{\sim} v_k$ et si la série $\sum v_k$ converge, alors la série $\sum u_k$ converge.*

b) *Supposons que $\sqrt[k]{u_k}$ tend vers une limite finie ℓ. Si $\ell < 1$, la série $\sum u_k$ converge ; si $\ell > 1$, la série ne converge pas.*

c) *Supposons que $\frac{u_{k+1}}{u_k}$ tend vers une limite finie ℓ. Si $\ell < 1$, la série $\sum u_k$ converge ; si $\ell > 1$, la série ne converge pas.*

Montrons le premier critère. Supposons $\lim\,(u_k/v_k) = 1$. Choisissons un nombre $\ell' > 1$. Pour tout k assez grand, on a $u_k/v_k \leqslant \ell'$, donc $u_k \leqslant \ell' v_k$ car v_k est positif pour k assez grand. Si la série $\sum v_k$ converge, il en va de même de la série $\sum (\ell' v_k)$, donc, par le théorème de comparaison, la série $\sum u_k$ converge.
Pour le deuxième critère, supposons $\ell < 1$ et choisissons un nombre ℓ' tel que $\ell < \ell' < 1$. Par hypothèse, on a $\sqrt[k]{u_k} \leqslant \ell'$ pour tout k assez grand, donc $u_k \leqslant \ell'^k$. La série géométrique $\sum \ell'^k$ a sa raison de module strictement inférieur à 1, donc est convergente. D'après le théorème de comparaison, la série $\sum u_k$ est donc convergente. Supposons maintenant $\ell > 1$. On a alors $\sqrt[k]{u_k} \geqslant 1$ pour tout k assez grand, donc aussi $u_k \geqslant 1$ et u_k ne tend pas vers 0 : la série n'est donc pas convergente.
Raisonnons de même pour le dernier critère en supposant d'abord $\ell < 1$. Pour tout k supérieur ou égal à un certain indice K, on a $\frac{u_{k+1}}{u_k} \leqslant \ell'$. En multipliant ces inégalités pour $k = K, K+1, \cdots, K+n$, on obtient $\frac{u_{K+n}}{u_K} \leqslant \ell'^n$, donc $u_{K+n} \leqslant u_K \ell'^n$ pour tout $n \geqslant 1$. Comme la série géométrique de raison ℓ' est convergente, il en va de même, par comparaison, de la série $\sum u_k$. Si $\ell > 1$, alors on a $u_{k+1} \geqslant u_k$ pour tout k assez grand, donc u_k ne tend pas vers 0.

Exemples

► La série $\sum \frac{z^k}{k!}$ est absolument convergente quel que soit le nombre complexe z.

En effet, si l'on pose $u_k = \dfrac{z^k}{k!}$, alors $\dfrac{|u_{k+1}|}{|u_k|} = \left|\dfrac{z^{k+1}}{z^k}\dfrac{k!}{(k+1)!}\right| = \dfrac{|z|}{k+1}$ tend vers 0 quand k tend vers $+\infty$. On en déduit que la série $\sum|u_k|$ est convergente, autrement dit : la série $\sum u_k$ est absolument convergente.

- Soit q un entier quelconque. Si z est un nombre complexe tel que $|z| < 1$, la série $\sum k^q z^k$ est absolument convergente.

En posant $u_k = k^q z^k$, le rapport $\dfrac{|u_{k+1}|}{|u_k|} = \left|\dfrac{k+1}{k}\right|^q |z|$ tend vers $|z|$ quand k tend vers $+\infty$. Si $|z|<1$, la série $\sum|u_k|$ est convergente, donc $\sum u_k$ est absolument convergente.

- La série $\sum \dfrac{k}{2k^2 - k + 1}$ n'est pas convergente, car son terme général est équivalent à $\dfrac{1}{2k}$ et la série de Riemann $\sum \dfrac{1}{k}$ n'est pas convergente.
- La série $\sum \dfrac{k}{\sqrt{k^5-1}}$ est convergente, car $\dfrac{k}{\sqrt{k^5-1}} \sim \dfrac{1}{k^{3/2}}$.
- Si P et Q sont des polynômes non nuls tels que $\deg Q - \deg P \geqslant 2$, la série $\sum \dfrac{P(k)}{Q(k)}$ est convergente.

En appelant p et q les degrés de P et Q, on a $P(k) \sim ak^p$ et $Q(k) \sim bk^q$, où a et b sont les coefficients dominant des polynômes. On en déduit $\dfrac{P(k)}{Q(k)} \sim \dfrac{a}{b}\dfrac{1}{k^{q-p}}$. La série de Riemann $\sum \dfrac{1}{k^{q-p}}$ est convergente si et seulement si $q-p > 1$, c'est-à-dire si et seulement si $q-p \geqslant 2$, puisque p et q sont des entiers.

2. Séries entières

Définition

Une série de la forme $\sum a_k z^k$, où z est un nombre réel ou complexe, s'appelle une *série entière.*

Les sommes partielles de la série entière $\sum a_k z^k$ sont les expressions polynomiales $s_n = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$, car par convention, on pose $z^0 = 1$.

Exemples

- La série $\sum \dfrac{z^k}{k!}$, de sommes partielles $s_n = 1 + z + \dfrac{z^2}{2!} + \cdots + \dfrac{z^n}{n!}$, est une série entière.
- De même, la série de sommes partielles $s_n = z + \dfrac{z^3}{2} + \dfrac{z^5}{3} + \cdots + \dfrac{z^{2n-1}}{n}$ est une série entière $\sum a_k z^k$: ses coefficients sont définis par $a_{2k} = 0$ pour $k \geqslant 0$ et $a_{2k-1} = 1/k$ pour tout $k \geqslant 1$.

Une série entière $\sum a_k z^k$ converge toujours pour $z = 0$. Considérons l'ensemble I des nombres réels $r \geqslant 0$ pour lesquels la série $\sum |a_k| r^k$ converge. Soit $r \in I$ et soit r' un nombre tel que $0 \leqslant r' \leqslant r$; la série $\sum |a_k| r^k$ converge et puisqu'on a $|a_k| r'^k \leqslant |a_k| r^k$, la série $\sum |a_k| r'^k$ converge aussi, donc $r' \in I$. Cela montre que si $r \in I$, alors I contient l'intervalle $[0, r]$. On en déduit que I est un intervalle de la forme $[0, R]$ ou bien $[0, R[$, R pouvant, dans ce dernier cas, être le symbole $+\infty$.

Supposons par exemple que R est un nombre réel. Alors pour tout nombre complexe z tel que $0 \leqslant |z| < R$, la série $\sum |a_k z^k| = \sum |a_k||z|^k$ est convergente, donc la série entière $\sum a_k z^k$ est absolument convergente.
Montrons que si $r > R$, la suite $a_k r^k$ n'est pas majorée. Raisonnons par l'absurde en supposant qu'il existe un nombre $M > 0$ tel que $|a_k|r^k \leqslant M$ pour tout k. Choisissons un nombre ρ tel que $R < \rho < r$. On a alors $|a_k|\rho^k = |a_k|r^k \left(\frac{\rho}{r}\right)^k \leqslant M \left(\frac{\rho}{r}\right)^k$ et comme $\frac{\rho}{r} < 1$, la série géométrique $\left(\frac{\rho}{r}\right)^k$ est convergente. Il s'ensuit que $\sum |a_k|\rho^k$ est convergente, ce qui est impossible puisque $\rho > R$.
On en déduit que si $r > R$, la suite $a_k z^k$ ne tend pas vers 0 : pour $r > R$, la série $\sum a_k z^k$ ne converge pas.

Théorème et définition. *Soit $\sum a_k z^k$ une série entière.*

- *Ou bien la série est absolument convergente pour tout z et l'on pose dans ce cas $R = \infty$,*
- *ou bien il existe un nombre réel $R \geqslant 0$ tel que la série est absolument convergente si $|z| < R$ et non convergente si $|z| > R$.*

R s'appelle le rayon de convergence de la série entière $\sum a_k z^k$.

Supposons que le rayon de convergence R est un nombre. Si $|z| > R$, la série ne converge pas, donc ne converge pas absolument. Le rayon est donc caractérisé par la propriété suivante : si $|z| < R$, il y a convergence absolue, et si $|z| > R$, il n'y a pas convergence absolue.

Si l'on a trouvé un nombre R tel que la série converge absolument pour $|z| < R$ et ne converge pas absolument pour $|z| > R$, alors le rayon de convergence est égal à R.

Calcul du rayon de convergence. *Soit $\sum a_k z^k$ une série entière dont les coefficients a_k sont non nuls.*

- *Si $\frac{|a_{k+1}|}{|a_k|}$ tend vers 0, alors le rayon de convergence est $R = \infty$.*
- *Si $\frac{|a_{k+1}|}{|a_k|}$ a une limite finie $\ell > 0$, alors le rayon de convergence est $\frac{1}{\ell}$.*

Démonstration. En posant $u_k = a_k z^k$, il vient $\frac{|u_{k+1}|}{|u_k|} = \frac{|a_{k+1}|}{|a_k|}|z|$. Supposons que $\frac{|a_{k+1}|}{|a_k|}$ a une limite finie ℓ, donc $\frac{|u_{k+1}|}{|u_k|}$ tend vers $\ell|z|$. Appliquons le critère de convergence (c). Si $\ell{=}0$, la série $\sum |u_k|$ converge quel que soit z, donc le rayon de convergence de $\sum a_k z^k$ est $R{=}\infty$. Si $\ell{>}0$, cette série converge pour $\ell|z|{<}1$, c'est-à-dire pour $|z|{<}1/\ell$, et elle ne converge pas pour $|z|{>}1/\ell$: d'après la règle ci-dessus, c'est donc que le rayon de convergence de $\sum a_k z^k$ est $R = 1/\ell$. ■

Exemples

- La série entière $\sum \frac{z^k}{k!}$ est absolument convergente pour tout z, donc son rayon de convergence est ∞.

► Pour la série entière $\sum \frac{2^k}{k} z^k$, on a $a_k = \frac{2^k}{k}$ et $\frac{|a_{k+1}|}{|a_k|} = 2\frac{k}{k+1}$ tend vers 2. Le rayon de convergence est donc $1/2$.

Pour $z = 1/2$, on obtient la série de Riemann $\sum \frac{1}{k}$ qui n'est pas convergente.

Pour $z = -1/2$, on obtient la série $\sum \frac{(-1)^k}{k}$ qui est une série alternée convergente, mais pas absolument convergente.

Ainsi, lorsque z parcourt l'ensemble des nombres complexes de module R, la nature de la série dépend de z.

► Considérons la série entière $\sum \frac{z^{2k}}{k^2+1}$. Puisque les coefficients des puissances impaires de z sont nuls, étudions la série de terme général $u_k = \frac{z^{2k}}{k^2+1}$. On a $\frac{|u_{k+1}|}{|u_k|} = \frac{k^2+1}{k^2+2k+2}|z|^2$, expression qui tend vers $|z|^2$. La série est donc absolument convergente si $|z| < 1$ et n'est pas absolument convergente si $|z| > 1$: on en déduit que le rayon de convergence est 1.

Convention : Si r est un nombre réel quelconque, nous convenons que l'on a $r < \infty$, donc aussi $\min\{r, \infty\} = r$ et $\max\{r, \infty\} = \infty$. On pose également $\min\{\infty, \infty\} = \infty = \max\{\infty, \infty\}$.

2.1 Opérations sur les séries entières

Série somme et série produit

étant données deux séries entières $\sum a_k z^k$ et $\sum b_k z^k$, la série $\sum (a_k+b_k)z^k$ est une série entière, ainsi que $\sum (\lambda a_k)z^k$ pour tout nombre λ.

Soient R et R' les rayons de convergence des séries entières $\sum a_k z^k$ et $\sum b_k z^k$. Si z est un nombre complexe tel que $|z| < R$ et $|z| < R'$, chacune des séries est absolument convergente, donc aussi leur somme $\sum (a_k+b_k)z^k$. On en déduit que le rayon de convergence de la série $\sum (a_k+b_k)z^k$ est au moins égal à $\min\{R, R'\}$. Si λ est un nombre non nul, les séries $\sum a_k z^k$ et $\sum (\lambda a_k)z^k$ convergent pour les mêmes valeurs de z, donc elles ont même rayon de convergence.

Définition
Soient $\sum a_k z^k$ et $\sum b_k z^k$ des séries entières. Leur *produit* est la série entière $\sum c_k z^k$, où $c_k = a_0 b_k + a_1 b_{k-1} + \cdots + a_{k-1} b_1 + a_k b_0$.

Le coefficient c_k est celui qui est en facteur de z^k dans le produit

$$(a_0+a_1 z+\cdots+a_k z^k)(b_0+b_1 z+\cdots+b_k z^k)$$

des k-ièmes sommes partielles. Pour calculer c_k, on développe donc ce produit en supprimant toutes les puissances de z supérieures à k, comme dans un développement limité.
Soient $\sum a_k z^k$ et $\sum b_k z^k$ des séries entières de rayon R et R'.

La série somme $\sum(a_k+b_k)z^k$ *et la série produit* $\sum c_k z^k$ *ont un rayon de convergence supérieur ou égal à* $\rho = \min\{R, R'\}$*. Pour* $|z| < \rho$*, on a*

$$\sum_{k=0}^{+\infty}(a_k+b_k)z^k = \sum_{k=0}^{+\infty} a_k z^k + \sum_{k=0}^{+\infty} b_k z^k \quad \text{et} \quad \sum_{k=0}^{+\infty} c_k z^k = \Big(\sum_{k=0}^{+\infty} a_k z^k\Big)\Big(\sum_{k=0}^{+\infty} b_k z^k\Big)$$

Propriétés de la somme d'une série entière

Soit $\sum a_k z^k$ une série entière de rayon de convergence $R > 0$. Sa somme $f(z)$ est définie si le module de z est strictement inférieur à R, autrement dit

> *la somme d'une série entière de rayon* $R > 0$ *est une fonction définie sur le disque ouvert* $D(O, R) = \{z \in \mathbb{C} \mid |z| < R\}$*, qu'on appelle le disque de convergence.*

Quand on s'en tient à des valeurs réelles de la variable, la somme de la série est définie dans l'intervalle ouvert $]-R, R[$.
Les propriétés de la somme $f(z)$ reposent sur l'existence d'une même « série majorante » lorsque $|z|$ décrit un segment $[-r, r]$ inclus dans $]-R, R[$:

Si r est un nombre tel que $0<r<R$, il existe une série géométrique convergente $\sum Mb^k$ où $0 < b < 1$, telle qu'on ait $|a_k z^k| \leqslant Mb^k$ pour tout k et pour tout z vérifiant $|z| \leqslant r$.

Choisissons en effet ρ tel que $r < \rho < R$. Puisque $\rho < R$, la suite $|a_k|\rho^k$ tend vers 0, donc on peut la majorer par un certain nombre M. Si $|z| \leqslant r$, alors on a

$$|a_k z^k| \leqslant |a_k| r^k = |a_k|\rho^k \left(\frac{r}{\rho}\right)^k \leqslant Mb^k \ , \ \text{avec } b = \frac{r}{\rho} < 1.$$

En utilisant cette propriété, montrons par exemple que la fonction f est continue en tout point z_0 du disque de convergence D.

Pour $z \in D$, on a $f(z)-f(z_0) = \sum_{k=0}^{+\infty} a_k(z^k - z_0^k)$. Puisque $|a_k||z^k - z_0^k| \leqslant |a_k||z|^k + |a_k||z_0|^k$, la série $\sum_{k=0}^{+\infty} |a_k|\,|z^k - z_0^k|$ est convergente, d'où

$$(1) \qquad \big|f(z)-f(z_0)\big| \leqslant \sum_{k=0}^{+\infty} |a_k|\,|z^k - z_0^k| = \sum_{k=0}^{K} |a_k|\,|z^k - z_0^k| + \sum_{k=K+1}^{+\infty} |a_k|\,|z^k - z_0^k|$$

Prenons r tel que $|z_0| < r < R$. Un nombre $\varepsilon > 0$ étant donné, commençons par choisir un entier K tel que $\sum_{k=K+1}^{+\infty} Mb^k \leqslant \varepsilon/2$, ce qui est possible puisque la série $\sum Mb^k$ converge. En utilisant à nouveau l'inégalité $|z^k - z_0^k| \leqslant |z|^k + |z_0|^k$, il vient pour $|z| < r$:

$$(2) \qquad \sum_{k=K+1}^{+\infty} |a_k|\,|z^k - z_0^k| \leqslant \sum_{k=K+1}^{+\infty} 2|a_k| r^k \leqslant \sum_{k=K+1}^{+\infty} 2Mb^k \leqslant \varepsilon$$

Les écarts $|z^k - z_0^k|$, pour $k = 0, 1, \ldots, K$, peuvent être rendus aussi petits qu'on veut en prenant z assez proche de z_0, car une fonction puissance est continue. La somme $\sum_{k=0}^{K} |a_k|\,|z^k - z_0^k|$ est donc aussi petite qu'on veut pourvu que $|z - z_0|$ soit assez petit et d'après (1) et (2), il en va de même de l'écart $\big|f(z)-f(z_0)\big|$.

Intégration et dérivation d'une série entière

Proposition. *Soit* $\sum a_k z^k$ *une série entière de rayon* $R>0$. *Pour tout* $x\in]-R,R[$, *on a*

$$\int_0^x \left(\sum_{k=0}^{+\infty} a_k t^k\right) dt = \sum_{k=0}^{+\infty} \frac{a_k}{k+1} x^{k+1} \quad \text{et} \quad \frac{d}{dx}\left(\sum_{k=0}^{+\infty} a_k x^k\right) = \sum_{k=1}^{+\infty} k a_k x^{k-1}$$

Pour intégrer ou dériver la somme d'une série entière sur l'intervalle de convergence, on dérive ou on intègre terme à terme, comme pour une somme ne comportant qu'un nombre fini de termes.

Démonstration. Montrons d'abord la formule d'intégration. Pour tout $t\in]-R,R[$, appelons $s_n(t)=a_0+a_1t+\cdots+a_nt^n$ les sommes partielles de la série et $f(t)=\lim s_n(t)$ la somme. Puisque la fonction f est continue, elle est intégrable entre 0 et $x\in]-R,R[$. Posons $F(x)=\int_0^x f(t)\,dt$ et $S_n(x)=\int_0^x s_n(t)\,dt=\sum_{k=0}^n \frac{a_k}{k+1}x^{k+1}$. On sait qu'il existe une série géométrique Mb^k convergente, avec $0<b<1$, telle que $|a_kt^k|\leqslant Mb^k$ pour tout k et pour tout t entre 0 et x. Il vient alors $\left|\frac{a_k}{k+1}t^{k+1}\right|\leqslant |t||a_kt^k|\leqslant RMb^k$ donc la série entière $\sum \frac{a_k}{k+1}t^{k+1}$ est absolument convergente pour t entre 0 et x. On en déduit

$$F(x)-S_n(x)=\int_0^x \left[f(t)-s_n(t)\right] dt = \int_0^x \left[\sum_{k=n+1}^{+\infty} \frac{a_k}{k+1}t^{k+1}\right] dt$$

Utilisons l'inégalité $\left|\int_0^x \varphi(t)\,dt\right| \leqslant \left|\int_0^x |\varphi(t)|\,dt\right|$ et que pour une série $\sum u_k$ absolument convergente, on a $\left|\sum_{k=0}^{+\infty} u_k\right| \leqslant \sum_{k=0}^{+\infty}|u_k|$ (page 531). Il vient

$$\left|F(x)-S_n(x)\right| \leqslant \left|\int_0^x \left(\sum_{k=n+1}^{+\infty}\left|\frac{a_k}{k+1}t^{k+1}\right|\right) dt\right| \leqslant \left|\int_0^x \left(\sum_{k=n+1}^{+\infty} RMb^k\right) dt\right|$$

En sommant la série géométrique, on obtient $\left|F(x)-S_n(x)\right|\leqslant\left|\int_0^x RM\frac{b^{n+1}}{1-b}\,dt\right| = \frac{RM|x|}{1-b}b^{n+1}$. Quand n tend vers l'infini, b^{n+1} tend vers 0 et donc $S_n(x)$ tend vers $F(x)$, ce qu'il fallait démontrer.

Pour montrer la formule de dérivation, il suffit de pouvoir appliquer la formule d'intégration à la série dérivée $\sum ka_kz^{k-1}$. Nous devons donc seulement montrer que son rayon de convergence est au moins R. Soit z un nombre complexe tel que $0<|z|<R$ et r tel que $0<|z|<r<R$. Il y a une série géométrique convergente Mb^k, avec $0<b<1$, telle que $|a_kz^k|\leqslant Mb^k$ pour tout k, donc $\left|ka_kz^k\right|\leqslant Mkb^k$. Pour la série de terme général $u_k=Mkb^k$, on a $\frac{u_{k+1}}{u_k}=\frac{k+1}{k}b$, expression qui tend vers $b<1$, donc $\sum u_k$ est convergente. On en déduit que la série $\sum ka_kz^k$ est absolument convergente, et en la multipliant par $1/z$, que la série $\sum ka_kz^{k-1}$ est convergente, d'où le résultat.

Si l'on note R_d le rayon de convergence de la série dérivée, cette démonstration prouve que l'on a $R_d\geqslant R$. Appliquons cela à la série intégrée, en notant R_i son rayon de convergence : en la dérivant, on obtient la série de départ, donc $R\geqslant R_i$. Mais comme la série intégrée converge pour $|z|<R$, on a aussi $R_i\geqslant R$, donc finalement $R_i=R$: en intégrant une série entière, on ne change pas son rayon de convergence ; donc en dérivant, non plus. ■

En dérivant ou en intégrant une série entière, on ne change pas son rayon de convergence.

Pour visualiser plus facilement la somme d'une série entière, notons-la sous la forme

$$\sum_{k=0}^{+\infty} a_k z^k = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n + \cdots$$

les trois points terminaux signifiant qu'on passe à la limite quand n tend vers $+\infty$.

Série entière et développement limité. Puisque la dérivée d'une série entière de rayon de convergence $R > 0$ est une série entière de même rayon, on peut dériver sa somme autant de fois qu'on veut sur l'intervalle $]-R, R[$.
Si $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \cdots$, alors

$$f'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + n a_n x^{n-1} + \cdots$$
$$f''(x) = 2a_2 + 3 \cdot 2\, x + 4 \cdot 3\, x^2 + \cdots + \cdots + n(n-1) x^{n-2} + \cdots$$

d'où $f(0) = a_0$, $f'(0) = a_1$, $f''(0) = 2a_2$ et $f^{(p)}(0) = p!\, a_p$, pour tout entier $p \geqslant 0$.
D'après la formule de Taylor-Young (page 301), on en déduit les résultats suivants.

- *Si une fonction f est la somme d'une série entière, cette série est unique.*
- *Si $f(x) = \sum_{k=0}^{+\infty} a_k x^k$, le développement limité de f à l'ordre n au point 0 est $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + o(x^n)$.*

2.2 Développement en série entière

Nous allons voir que les fonctions usuelles s'expriment généralement comme la somme d'une série entière, du moins au voisinage d'un point.

Principaux développements en série entière	
Fonction Développement	R
$\frac{1}{a-z} = \sum_{k=0}^{+\infty} \frac{z^k}{a^{k+1}} = \frac{1}{a} + \frac{z}{a^2} + \frac{z^2}{a^3} + \frac{z^3}{a^4} + \cdots + \frac{z^n}{a^{n+1}} + \cdots$	$\lvert a \rvert$
$\ln(1+x) = \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{k} x^k = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \cdots + \frac{(-1)^{n+1}}{n} x^n + \cdots$	1
$e^z = \sum_{k=0}^{+\infty} \frac{z^k}{k!} = 1 + z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \cdots + \frac{1}{n!}z^n + \cdots$	∞
$\sin x = \sum_{k=0}^{+\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \cdots + \frac{(-1)^n}{(2n+1)!} x^{2n+1} + \cdots$	∞
$\cos x = \sum_{k=0}^{+\infty} \frac{(-1)^k}{(2k)!} x^{2k} = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 + \cdots + \frac{(-1)^n}{(2n)!} x^{2n} + \cdots$	∞
$\mathrm{Arctan}\, x = \sum_{k=1}^{+\infty} \frac{(-1)^{k+1}}{2k+1} x^{2k+1} = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 + \cdots + \frac{(-1)^{n+1}}{2n+1} x^{2n+1} + \cdots$	1
$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2!} x^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3!} x^3 + \cdots + \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^n + \cdots$	1

Pour démontrer ces formules, on pratique la dérivation et l'intégration, comme on l'a fait au chapitre 10 pour calculer les développements limités usuels.

1) Le développement de $\frac{1}{a-z} = \frac{1}{a}\frac{1}{1-(z/a)}$ est simplement le produit par $\frac{1}{a}$ de la somme d'une série géométrique dont la raison z/a est de module $|z/a|<1$. En intégrant pour x réel le développement

$$\frac{1}{1+x} = 1-x+x^2-x^3+\cdots+(-1)^n x^n+\cdots \quad \text{de rayon } 1,$$

on trouve celui de $\ln(1+x)$ qui a donc même rayon de convergence $R=1$.

2) Soit a un nombre complexe. La série entière $\sum \frac{a^k}{k!}x^k = \sum \frac{(ax)^k}{k!}$ est convergente pour tout réel x (exemple page 532), donc son rayon de convergence est ∞. Posons $f(x) = \sum_{k=0}^{+\infty} \frac{a^k}{k!}x^k = 1+ax+\frac{a^2}{2}x^2+\cdots+\frac{a^n}{n!}x^n+\cdots$. En dérivant, il vient

$$\begin{aligned} f'(x) &= a+a^2x+\frac{a^3}{2}x^2+\cdots+\frac{a^n}{(n-1)!}x^{n-1}+\cdots \\ &= a\left[1+ax+\frac{a^2}{2}x+\cdots+\frac{a^{n-1}}{(n-1)!}x^{n-1}+\cdots\right] = af(x) \end{aligned}$$

donc f est solution de l'équation différentielle $y'=ay$. Or les solutions de cette équation sont les fonctions λe^{ax} (remarque page 444). Puisqu'on a $f(0)=1$, il vient $f(x)=e^{ax}$. En particulier, on a $\sum_{k=0}^{+\infty}\frac{a^k}{k!} = f(1) = e^a$, et cela quel que soit a complexe. Il suffit de substituer z à a pour avoir le développement en série entière de la fonction e^z formulé dans le tableau.

3) En prenant $a=\boldsymbol{i}$ puis $a=-\boldsymbol{i}$, on obtient pour tout x réel

$$\begin{aligned} e^{\boldsymbol{i}x} &= \sum_{k=0}^{+\infty}\frac{\boldsymbol{i}^k}{k!}x^k = 1+\boldsymbol{i}x-\frac{1}{2}x^2-\frac{\boldsymbol{i}}{3!}x^3+\frac{1}{4!}x^4+\cdots+\frac{\boldsymbol{i}^n}{n!}x^n+\cdots \\ e^{-\boldsymbol{i}x} &= \sum_{k=0}^{+\infty}\frac{(-1)^k\boldsymbol{i}^k}{k!}x^k = 1-\boldsymbol{i}x-\frac{1}{2}x^2+\frac{\boldsymbol{i}}{3!}x^3+\frac{1}{4!}x^4+\cdots+\frac{(-1)^n\boldsymbol{i}^n}{n!}x^n+\cdots \end{aligned}$$

Ajoutons ces deux séries entières : les puissances impaires de x s'annulent et les puissances paires sont les mêmes, donc il reste $e^{\boldsymbol{i}x}+e^{-\boldsymbol{i}x} = 2\left[1-\frac{1}{2}x^2+\frac{1}{4!}x^4+\cdots+\frac{(-1)^n}{(2n)!}x^{2n}+\cdots\right]$ et l'on trouve ainsi le développement en série entière de $\cos x$.

Pour avoir celui de $\sin x$, il suffit de soustraire les deux séries entières.

4) La dérivée de $\operatorname{Arc}\tan x$ est $\frac{1}{1+x^2}$ et pour $|x|<1$, on a le développement en série entière $\frac{1}{1+x^2} = 1-x^2+x^4+\cdots+(-1)^n x^{2n}+\cdots$ en remplaçant x par x^2 dans le développement de $\frac{1}{1+x}$. Intégrons terme à terme et tenons compte de $\operatorname{Arc}\tan(0)=0$: on trouve le développement de $\operatorname{Arc}\tan x$ et le rayon de convergence est encore 1.

5) Nous calculerons le développement de $(1+x)^\alpha$ dans l'exemple page 541.

Exemple 1. Calculons la somme $f(x)=\sum_{k=0}^{+\infty}\frac{x^{2k}}{2k+1}$.

La série est absolument convergente si $|x|<1$, car on a alors $\left|\frac{x^{2k}}{2k+1}\right|\leqslant|x|^{2k}$ et la série géométrique $\sum|x^2|^k$ a pour raison $|x|^2<1$. Si $x>1$, le terme général $\frac{x^{2k}}{2k+1}$ tend vers l'infini, donc la série ne converge pas.

Le rayon de convergence est donc 1 et la fonction f est définie sur $]-1,1[$.

Posons $g(x)=xf(x)$, ou encore

$$g(x)=\sum_{k=0}^{+\infty}\frac{x^{2k+1}}{2k+1}=x+\frac{x^3}{3}+\frac{x^5}{5}+\cdots+\frac{x^{2n+1}}{2n+1}+\cdots$$

En dérivant, il vient

$$g'(x)=1+x^2+x^4+\cdots+x^{2n}+\cdots=\frac{1}{1-x^2}=\frac{1}{2}\left[\frac{1}{1+x}+\frac{1}{1-x}\right]$$

donc $g(x)=\frac{1}{2}\left[\ln(1+x)-\ln(1-x)\right]$ et $f(x)=\frac{1}{2x}\ln\frac{1+x}{1-x}$. Cette expression est bien définie en $x=0$: en effet, $\left(\ln\frac{1+x}{1-x}\right)\underset{x\to0}{\sim}2x$, donc $\lim_{x\to0}\left(\frac{1}{2x}\ln\frac{1+x}{1-x}\right)=1=f(0)$.

Exemple 2. Développons en série entière la fonction $g(x)=\int_0^x\frac{1}{t}\sin(at^2)\,dt$.

Quand t tend vers 0, $\frac{1}{t}\sin(at^2)\sim\frac{at^2}{t}=at$, donc la fonction sous le signe intégrale est continue en 0. Son développement est

$$\begin{aligned}\frac{1}{t}\sin(at^2)&=\frac{1}{t}\left[at^2-\frac{(at^2)^3}{3!}+\cdots+(-1)^n\frac{(at^2)^{2n+1}}{(2n+1)!}+\cdots\right]\\&=at-\frac{a^3t^5}{3!}+\cdots+(-1)^n\frac{a^{2n+1}t^{4n+1}}{(2n+1)!}+\cdots\end{aligned}$$

et le rayon de convergence est ∞. D'où en intégrant de 0 à x :

$$\begin{aligned}g(x)&=a\frac{x^2}{2}-\frac{a^3}{3!}\frac{x^6}{6}+\cdots+(-1)^n\frac{a^{2n+1}}{(2n+1)!}\frac{x^{4n+2}}{4n+2}+\cdots\\&=\frac{ax^2}{2}\left[1-\frac{a^2}{3!}\frac{x^4}{3}+\cdots+(-1)^n\frac{a^{2n}}{(2n+1)!}\frac{x^{4n}}{2n+1}+\cdots\right]\end{aligned}$$

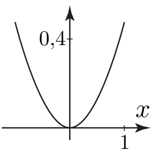
avec un rayon de convergence ∞.

Comme la série est alternée, on sait que si l'on remplace sa somme par une somme partielle, on fait une erreur moindre que le premier terme négligé.

Ainsi dans le cas $a=1$, si l'on pose $\sigma(x)=\frac{x^2}{2}-\frac{x^6}{6(3!)}$, alors

$$0\leqslant g(x)-\sigma(x)\leqslant\frac{|x|^{10}}{1200}$$

Entre -1 et 1, l'erreur commise en remplaçant $g(x)$ par le polynôme $\sigma(x)$ est moindre que 10^{-3}.



graphe de $\sigma(x)$

2.3 Calculs de solutions d'équations différentielles

De nombreuses solutions d'équations différentielles peuvent s'exprimer sous la forme d'une série entière. Si la fonction inconnue s'appelle y, on fait la substitution $y=\sum a_k x^k$ dans l'équation différentielle et l'on identifie les coefficients de chaque puissance de x.

Exemple. Posons $f(x)=(1+x)^\alpha$. Puisque $f'(x)=\alpha(1+x)^{\alpha-1}=\frac{\alpha}{1+x}f(x)$ la fonction f est solution de l'équation différentielle

$$(*) \qquad (1+x)y' = \alpha y$$

C'est la seule solution telle que $f(0)=1$. Posons $y=\sum a_k x^k$ et remplaçons dans $(*)$:

- dans αy, le coefficient de x^k est αa_k,
- dans y', le coefficient de x^k est $(k+1)a_{k+1}$,
- dans xy', le coefficient de x^k est $k a_k$,

donc il vient $(k+1)a_{k+1} + k a_k = \alpha a_k$ pour tout k. Pour $k = 0, 1, \ldots, n$, on obtient

$$\begin{aligned} a_1 &= \alpha a_0 \\ 2a_2 &= (\alpha-1)a_1 \\ \ldots & \quad \ldots \\ n a_n &= (\alpha-n+1)a_{n-1}\ , \quad \text{d'où en multipliant membre à membre} \\ (n!)a_n &= \alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)a_0 \quad \text{et} \quad a_n = \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}a_0\,. \end{aligned}$$

Puisque $a_0 = f(0) = 1$, on en déduit le développement

$$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2!}x^2 + \cdots + \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}x^n + \cdots$$

Si α est un entier $p \geqslant 0$, le coefficient a_{p+1} est nul, donc $a_k = 0$ quel que soit $k > p$: dans ce cas, selon la formule du binôme, la série est simplement le polynôme

$$(1+x)^p = 1+\binom{p}{1}x+\binom{p}{2}x^2+\cdots+\binom{p}{k}x^k+\cdots+\binom{p}{p}x^p\,.$$

Dans le cas général, aucun coefficient a_k n'est nul et $\left|\frac{a_n}{a_{n-1}}\right| = \frac{|\alpha-n+1|}{n}$ tend vers 1 quand n tend vers $+\infty$, donc le rayon de convergence est 1.

Voici les cas particuliers de cette formule pour $\alpha = -1/2$ et pour $\alpha = 1/2$.

$\frac{1}{\sqrt{1-x}} = 1 + \frac{1}{2}x + \frac{1\cdot 3}{2\cdot 4}x^2+\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}x^3 + \cdots + \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}x^n + \cdots$	$R=1$
$\sqrt{1-x} = 1 - \frac{1}{2}x - \frac{1}{2\cdot 4}x^2 - \frac{1\cdot 3}{2\cdot 4\cdot 6}x^3 - \cdots - \frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots 2n}x^n + \cdots$	$R=1$

Les fonctions de Bessel

L'équation différentielle de Bessel est

$$(b_\alpha) \qquad x^2y'' + xy' + (x^2 - \alpha^2)y = 0$$

où α est un nombre réel (exercice 6 page 474).

Cherchons une solution (non nulle) sous la forme $y(x) = x^\alpha \sum_{k=0}^{+\infty} a_k x^k = \sum_{k=0}^{+\infty} a_k x^{k+\alpha}$.

Dérivons pour $x > 0$:

$$y' = \alpha x^{\alpha-1} \sum_{k=0}^{+\infty} a_k x^k + x^\alpha \sum_{k=0}^{+\infty} k a_k x^{k-1} = \sum_{k=0}^{+\infty} (\alpha+k) a_k x^{\alpha+k-1}$$

$$y'' = \sum_{k=0}^{+\infty} (\alpha+k)(\alpha+k-1) a_k x^{\alpha+k-2}$$

- Dans x^2y, le coefficient de $x^{\alpha+k}$ est a_{k-2} pour $k \geqslant 2$ et 0 sinon,
- dans $\alpha^2 y$, le coefficient de $x^{\alpha+k}$ est $\alpha^2 a_k$,
- dans xy', le coefficient de $x^{\alpha+k}$ est $(\alpha+k)a_k$,
- et dans x^2y'', le coefficient de $x^{k+\alpha}$ est $(\alpha+k)(\alpha+k-1)a_k$.

En écrivant que dans le premier membre de l'équation, le coefficient de $x^{k+\alpha}$ doit être nul, on obtient $(\alpha+k)(\alpha+k-1)a_k + (\alpha+k)a_k + a_{k-2} - \alpha^2 a_k = 0$ si $k \geqslant 2$. Il vient

$$(2\alpha+1)a_1 = 0 \quad \text{et} \quad a_{k-2} + k(2\alpha+k)a_k = 0 \quad \text{pour tout } k \geqslant 2.$$

Cas où 2α n'est pas entier. On a alors $2\alpha+k \neq 0$ pour tout k, donc aussi $a_1 = 0$ d'après la première relation. Pour $k = 3, 5, \ldots$, la seconde relation donne alors $a_3 = a_5 = \cdots = 0$: les coefficients d'indices impairs sont tous nuls. Pour les coefficients d'indices pairs, il vient $\dfrac{a_{2k}}{a_{2k-2}} = \dfrac{-1}{2k(2\alpha+2k)}$, donc

$$a_{2k} = \frac{(-1)^k a_0}{2^{2k} k!(\alpha+1)(\alpha+2)\cdots(\alpha+k)}, \quad \text{pour tout } k \geqslant 1.$$

Puisque $\dfrac{a_{2k}}{a_{2k-2}}$ tend vers 0 quand k tend vers $+\infty$, le rayon de convergence de la série $\sum a_k x^k$ est ∞. Finalement, en choisissant $a_0 = 1$, la fonction

$$J_\alpha(x) = x^\alpha \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!(\alpha+1)(\alpha+2)\cdots(\alpha+k)} \left(\frac{x}{2}\right)^{2k}$$

est solution de l'équation de Bessel (b_α) sur l'intervalle $]0, +\infty[$.

Supposons par exemple $\alpha > 0$. Comme les équations (b_α) et $(b_{-\alpha})$ sont les mêmes, on trouve que si 2α n'est pas un entier, les fonctions J_α et $J_{-\alpha}$ sont solutions. Mais J_α est définie en 0, alors que $J_{-\alpha}$ ne l'est pas. Ces fonctions ne sont donc pas proportionnelles : par conséquent, elles forment une base de solutions de (b_α), autrement dit :

si 2α n'est pas entier, les solutions de l'équation de Bessel (b_α) sur $]0, +\infty[$ sont les combinaisons $\lambda J_\alpha + \mu J_{-\alpha}$, où λ et μ sont des constantes quelconques.

Cas où α est un entier p. Si $p \geqslant 0$, on obtient de même la solution

$$J_p(x) = \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!(p+1)\cdots(p+k)} \left(\frac{x}{2}\right)^{p+2k}$$

qui est cette fois la somme d'une série entière de rayon de convergence ∞.

Pour chercher une autre solution, reprenons les calculs précédents avec $\alpha = -p$ et $p \geqslant 1$: les relations s'écrivent $(-2p+1)a_1 = 0$ et $a_{k-2} + k(-2p+k)a_k = 0$ pour tout $k \geqslant 2$. On en tire $a_1 = 0$ et comme $-2p+k \neq 0$ pour k impair, il vient encore successivement $a_1 = a_3 = \cdots = a_{2i+1} = 0$ pour tout $i \geqslant 0$. En prenant $k = 2p$, la seconde relation donne $a_{2p-2} = 0$, et par suite $a_{2p-2} = a_{2p-4} = \cdots = a_0 = 0$. Pour $2k > 2p$, on a $a_{2k} = \dfrac{-a_{2k-2}}{2k(2k-2p)}$ et pour $i \geqslant 0$, les coefficients d'indice $2p+2i$ sont déterminés par a_{2p} : on a $a_{2p+2i} = \dfrac{(-1)^i}{2^{2i}(p+1)\cdots(p+i)i!} a_{2p}$ pour tout $i \geqslant 0$. Puisqu'on a pris $\alpha = -p$, la solution obtenue est

$$a_{2p}x^{-p}\sum_{i=0}^{+\infty} \frac{(-1)^i}{2^{2i}(p+1)\cdots(p+i)i!} x^{2p+2i} = 2^p a_{2p} \sum_{i=0}^{+\infty} \frac{(-1)^i}{(p+1)\cdots(p+i)i!} \left(\frac{x}{2}\right)^{p+2i} = 2^p a_{2p} J_p(x)$$

Ainsi, on trouve une solution proportionnelle à J_p : la méthode ne donne pas toutes les solutions de l'équation différentielle.

La fonction J_0. Rappelons que la fonction $x \mapsto \dfrac{1}{\pi}\displaystyle\int_0^\pi \cos(x \sin\theta)\, d\theta$ est une solution de l'équation de Bessel (b_0) qui prend la valeur 1 en $x = 0$ (exercice 8 page 394) et que les seules solutions définies en $x = 0$ sont proportionnelles à cette fonction (exercice 4 page 473). La série entière J_0 prend la valeur 1 en $x = 0$: c'est donc que l'on a pour tout nombre réel x l'égalité

$$\frac{1}{\pi}\int_0^\pi \cos(x \sin\theta)\, d\theta = J_0(x) = \sum_{k=0}^{+\infty} \frac{(-1)^k}{(k!)^2}\left(\frac{x}{2}\right)^{2k} = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2\,4^2} - \frac{x^6}{2^2\,4^2\,6^2} + \cdots$$

Remarquons que la série est alternée. En dérivant terme à terme, on vérifie que $J_0'(x) = -J_1(x)$.

2.4 Polynômes de Legendre

Étant donné un entier $n \geqslant 1$, considérons *l'équation différentielle de Legendre*

(L_n) $$(1 - x^2)y'' - 2xy' + n(n+1)y = 0$$

Cherchons des solutions sous forme d'une série entière $y = \sum a_k x^k$.

- Dans $2xy'$, le coefficient de x^k est $2ka_k$,
- dans x^2y'', le coefficient de x^k est $k(k-1)a_k$,
- et dans y'', le coefficient de x^k est $(k+2)(k+1)a_{k+2}$,

donc on a pour tout k la relation $(k+2)(k+1)a_{k+2} - k(k-1)a_k - 2ka_k + n(n+1)a_k = 0$, ou encore

$(*)$ $$(k+2)(k+1)a_{k+2} = -(n-k)(n+k+1)a_k$$

Puisque n est entier, le terme de droite est nul pour $k=n$, donc tous les coefficients a_{n+2i} sont nuls pour $i \geqslant 1$.

Supposons n pair, $n = 2m$

a) Supposons $a_0 = 0$. Alors d'après la relation $(*)$, tous les coefficients d'indices pairs sont nuls. Le coefficient a_1 détermine $a_3 = \dfrac{-(n-1)(n+2)}{2\cdot 3}a_1$, $a_5 = \dfrac{(n-3)(n-1)(n+2)(n+4)}{2\cdot 3\cdot 4\cdot 5}a_1$ et plus généralement, on a

$$a_{2k+1} = (-1)^k \frac{(n-2k+1)(n-2k-1)\cdots(n-1)(n+2)(n+4)\cdots(n+2k)}{(2k+1)!}a_1$$

On obtient ainsi une série entière $y = \sum_{k=0}^{\infty} a_{2k+1}x^{2k+1}$ solution de l'équation (L_n). Si $a_1 \neq 0$, le rayon de convergence est 1, car d'après $(*)$, $\dfrac{|a_{k+2}|}{|a_k|}$ tend vers 1 quand k tend vers $+\infty$.

b) Supposons $a_1 = 0$. Cette fois, tous les coefficients d'indices impairs sont nuls, de même que tous les coefficients d'indices pairs supérieurs à $n = 2m$: on va donc trouver un polynôme pair. La relation $(*)$ donne $(n-k+2)(n+k-1)a_{k-2} = -k(k-1)a_k$ et le coefficient $a_n = a_{2m}$ détermine $a_{2m-2}, a_{2m-4}, \ldots, a_0$, selon la formule

$$a_{2m-2k} = (-1)^k \frac{n!(n-1)!}{2\cdot(2n-1)!}\,\frac{(2n-2k)!}{k!(n-k)!(n-2k)!}a_{2m}, \text{ pour } 0 \leqslant k \leqslant m.$$

En choisissant $a_{2m} = a_n = \dfrac{(2n)!}{2^n(n!)^2}$, on obtient comme solution de l'équation différentielle (L_n) le polynôme

$$P_n(x) = \sum_{k=0}^{m} (-1)^k \frac{(2n-2k)!}{2^n k!(n-k)!(n-2k)!}x^{n-2k}$$

Cas où n est impair. En posant $n = 2m+1$, la même formule que ci-dessus définit un polynôme P_n solution de l'équation différentielle (L_n).

Quel que soit n, le polynôme P_n est de degré n et a la parité de n. Les polynômes P_n s'appellent les *polynômes de Legendre*. Ainsi

$$P_0 = 1\ ,\ P_1 = x\ ,\ P_2 = \frac{1}{2}(3x^2-1)\ ,\ P_3 = \frac{1}{2}(5x^3-3x)\ ,\ P_4 = \frac{1}{8}(35x^4-30x^2+3)$$

On peut calculer de proche en proche les polynômes de Legendre au moyen de la relation suivante :

$$(n+2)P_{n+2} = (2n+3)xP_{n+1} - (n+1)P_n \text{ , pour tout entier } n \geqslant 0.$$

Ces polynômes interviennent souvent en Analyse. En voici quelques propriétés.

Formule de Rodrigues : $P_n(x) = \dfrac{1}{2^n\cdot n!}\dfrac{d^n}{dx^n}\left[(x^2-1)^n\right]$

D'après la formule du binôme, $(x^2-1)^n = \sum_{k=0}^n (-1)^k \dfrac{n!}{k!(n-k)!} x^{2n-2k}$. Pour tout $q \geqslant n$, la dérivée n-ième de x^q est $q(q-1)\cdots(q-n+1)x^{q-n} = \dfrac{q!}{(q-n)!}x^{q-n}$. Il vient donc $\dfrac{d^n}{dx^n}x^{2n-2k} = \dfrac{(2n-2k)!}{(n-2k)!}x^{n-2k}$ pour $2k \leqslant n$, c'est-à-dire pour $k \leqslant m$ puisque m vaut $\dfrac{n}{2}$ ou $\dfrac{n-1}{2}$ selon la parité de n. Finalement, on obtient

$$\frac{d^n}{dx^n}\left[(x^2-1)^n\right] = \sum_{k=0}^{m} (-1)^k \frac{n!}{k!(n-k)!}\,\frac{(2n-2k)!}{(n-2k)!}x^{n-2k} = 2^n \cdot n!\, P_n(x)\,.$$

Relations d'orthogonalité : $\displaystyle\int_{-1}^{1} P_n(t)P_q(t)\,dt = \begin{cases} 0 & \text{si } n \neq q \\ \dfrac{2}{2n+1} & \text{si } n = q \end{cases}$

Si $r < q$, la dérivée $\dfrac{d^r}{dx^r}\left[(x^2-1)^q\right]$ est divisible par $(x^2-1)^{q-r}$, donc s'annule en $x = \pm 1$. Supposons $q \geqslant n$. En intégrant successivement par parties, il vient

$$\begin{aligned}
&\int_{-1}^{1} \frac{d^q}{dx^q}\left[(x^2-1)^q\right]\frac{d^n}{dx^n}\left[(x^2-1)^n\right]dx \\
&= \left[\frac{d^{q-1}}{dx^{q-1}}\left[(x^2-1)^q\right]\frac{d^n}{dx^n}\left[(x^2-1)^n\right]\right]_{-1}^{1} - \int_{-1}^{1}\frac{d^{q-1}}{dx^{q-1}}\left[(x^2-1)^q\right]\frac{d^{n+1}}{dx^{n+1}}\left[(x^2-1)^n\right]dx \\
&= \qquad \cdots\cdots \\
&= (-1)^q \int_{-1}^{1} (x^2-1)^q \frac{d^{q+n}}{dx^{q+n}}\left[(x^2-1)^n\right]dx\ , \quad \text{car les crochets sont nuls.}
\end{aligned}$$

Si $q > n$, alors on a $q+n > 2n$ et comme le polynôme $(x^2-1)^n$ est de degré $2n$, sa dérivée $(q+n)$-ième est nulle : on donc la formule annoncée pour $q > n$.

Supposons $q = n$. D'après le calcul ci-dessus, il vient

$$\begin{aligned}
\int_{-1}^{1} \frac{d^n}{dx^n}\left[(x^2-1)^n\right]\frac{d^n}{dx^n}\left[(x^2-1)^n\right]dx &= (-1)^n \int_{-1}^{1}(x^2-1)^n\frac{d^{2n}}{dx^{2n}}\left[(x^2-1)^n\right]dx \\
&= (2n)!\int_{-1}^{1}(1-x^2)^n\,dx\ , \quad \text{car } \frac{d^{2n}}{dx^{2n}}\left[(x^2-1)^n\right] = (2n)! \\
&= 2\cdot(2n)!\int_{0}^{1}(1-x^2)^n\,dx\ , \quad \text{car la fonction à intégrer est paire} \\
&= 2\cdot(2n)!\int_{0}^{\pi/2}(\sin\theta)^{2n+1}\,d\theta\ , \quad \text{en posant } x = \cos\theta \\
&= 2\cdot(2n)!\,\frac{2\cdot 4\cdots(2n)}{3\cdot 5\cdots(2n+1)} \quad \text{(exercice 13 page 341).}
\end{aligned}$$

On a donc $\displaystyle\int_{-1}^{1}\left[P_n(x)\right]^2 dx = \frac{2\cdot(2n)!}{(2^n\cdot n!)^2}\,\frac{2\cdot 4\cdots(2n)}{3\cdot 5\cdots(2n+1)} = \frac{2}{2n+1}$, car $2^n\cdot n! = 2\cdot 4\cdots(2n)$.

Un espace euclidien de fonctions. Pour toutes fonctions f et g continues sur $[-1,1]$, posons

$$f \cdot g = \int_{-1}^{1} f(x)g(x)\,dx$$

On a $f\cdot g = g\cdot f$ et puisque intégrer est une opération linéaire, on a $(f_1+f_2)\cdot g = f_1\cdot g + f_2\cdot g$ et $(\lambda f)\cdot g = \lambda(f\cdot g)$: le nombre $f\cdot g$ dépend donc linéairement de chaque terme. De plus, $f\cdot f = \int_{-1}^{1}\big[f(x)\big]^2\,dx$ est strictement positif si f n'est pas identiquement nulle sur $[-1,1]$.

On a donc ainsi défini un produit scalaire sur l'espace vectoriel V des fonctions continues sur $[-1,1]$ (chapitre 7).
D'après les relations précédentes, les polynômes de Legendre $P_0, P_1, \ldots, P_k \ldots$ sont deux à deux orthogonaux pour ce produit scalaire. Puisque P_k est de degré k, les $(n+1)$ polynômes $P_0, P_1, \ldots, P_n$ forment une base du sous-espace $\mathbb{P}_n$ des polynômes de degré inférieur ou égal à n. En divisant chacun par sa norme, on obtient une base orthonormée de $\mathbb{P}_n$.

Les polynômes $\sqrt{1/2}\,P_0, \sqrt{3/2}\,P_1, \ldots, \sqrt{(2n+1)/2}\,P_n$ forment une base orthonormée de l'espace $\mathbb{P}_n$ muni du produit scalaire $P\cdot Q = \int_{-1}^{1} P(x)Q(x)\,dx$.

Soit f une fonction appartenant à V. Parmi les polynômes P de degré au plus n, celui qui rend minimum la distance $\|f-P\|$ est, d'après le théorème de la projection page 209, le projeté orthogonal de f sur le sous-espace $\mathbb{P}_n$: son expression est $P = \frac{1}{2}(f\cdot P_0)P_0 + \frac{3}{2}(f\cdot P_1)P_1 + \cdots + \frac{2n+1}{2}(f\cdot P_n)P_n$.

Des harmoniques sphériques

Repérons les points M de $\mathbb{R}^3$ par des coordonnées sphériques

$$x = r\cos\theta\sin\psi \quad , \quad y = r\sin\theta\sin\psi \quad , \quad z = r\cos\psi$$

où r est la distance OM, $\theta = \widehat{\overrightarrow{Ox}, \overrightarrow{Om}}$ (m est le projeté de M sur xOy) et $\psi = \widehat{\overrightarrow{Oz}, \overrightarrow{OM}}$ $(0 \leqslant \theta < 2\pi,\ 0 \leqslant \psi \leqslant \pi)$.
Le laplacien d'une fonction $\Phi = r^n S(\theta,\psi)$ est

$$\Delta\Phi = r^{n-2}\left[\frac{1}{\sin\psi}\frac{\partial}{\partial\psi}\big[(\sin\psi)\frac{\partial S}{\partial\psi}\big] + \frac{1}{\sin^2\psi}\frac{\partial^2 S}{\partial\theta^2} + n(n+1)S\right]$$

Supposons que les valeurs de Φ présentent une symétrie de révolution par rapport à l'axe Oz : cela veut dire que S ne dépend que de l'angle ψ. La condition $\Delta\Phi = 0$ pour que $\Phi = r^n S(\psi)$ soit harmonique est donc

$$(H) \qquad \frac{d}{d\psi}\big[(\sin\psi)\frac{dS}{d\psi}\big] + n(n+1)(\sin\psi)S = 0$$

Posons $\mu = \cos\psi$. Il vient $\frac{dS}{d\psi} = \frac{dS}{d\mu}\frac{d\mu}{d\psi} = -\frac{dS}{d\mu}\sin\psi$, donc $(\sin\psi)\frac{dS}{d\psi} = (\mu^2-1)\frac{dS}{d\mu}$ et

$$\frac{d}{d\psi}\big[(\sin\psi)\frac{dS}{d\psi}\big] = \frac{d}{d\mu}\left[(\mu^2-1)\frac{dS}{d\mu}\right]\frac{d\mu}{d\psi} = (\sin\psi)\frac{d}{d\mu}\left[(1-\mu^2)\frac{dS}{d\mu}\right]$$

La condition (H) s'écrit $\dfrac{d}{d\mu}\left[(1-\mu^2)\dfrac{dS}{d\mu}\right] + n(n+1)S = 0$, ou encore

$$(h) \qquad (1-\mu^2)\frac{d^2S}{d\mu^2} - 2\mu\frac{dS}{d\mu} + n(n+1)S = 0$$

Pour n entier positif, c'est l'équation différentielle de Legendre (L_n). Si n est négatif, alors en posant $k = -n-1 > 0$, on a $n(n+1) = k(k+1)$, donc $S(\mu)$ doit être solution de L_k. On a ainsi le résultat suivant.

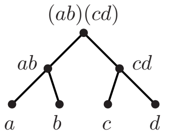
Pour tout entier n positif, les fonctions $r^n P_n(\cos\psi)$ et $\dfrac{1}{r^{n+1}} P_n(\cos\psi)$ sont harmoniques.

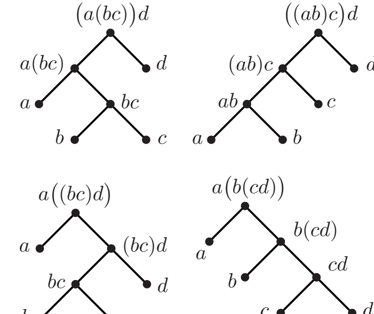
► Donnons-nous sur la sphère de rayon a, une fonction $f(\psi)$ combinaison de polynômes de Legendre en $\cos\psi$: $f(\psi) = c_0P_0(\cos\psi)+c_1P_1(\cos\psi)+\cdots+c_mP_m(\cos\psi)$. Alors la fonction $F(r,\psi) = c_0P_0(\cos\psi) + c_1\frac{r}{a}P_1(\cos\psi) + \cdots + c_m\frac{r^m}{a^m}P_m(\cos\psi)$ est harmonique dans la boule de rayon a et vaut f sur le bord de la boule : en effet, pour $r = a$, on a $F(a,\psi) = f(\psi)$. Si l'on cherche dans la boule un potentiel harmonique qui satisfait à peu près une condition $c(\psi)$ donnée sur le bord, on peut commencer par approcher $c(\psi)$ par une combinaison $f(\psi)$ de polynômes de Legendre, comme en fin de paragraphe précédent, et utiliser $F(r,\psi)$.

► Sur une sphère, chaque fonction $\Phi_n = r^nP_n(\cos\psi)$ ou $\dfrac{1}{r^{n+1}}P_n(\cos\psi)$ s'annule le long des parallèles ψ=constante telle que $\cos\psi$ est une racine du polynôme P_n ; cela découpe la sphère en bandes horizontales appelées *zones* : les Φ_n s'appellent des *harmoniques zonales*.

2.5 Un exemple de fonction génératrice

Considérons les arbres ayant un seul sommet initial et exactement deux arêtes issues de chaque sommet non terminal : ce sont les *arbres binaires*.





On les utilise en algorithmique pour décrire les différentes façons de parcourir un processus où il y a deux possibilités à chaque étape. Ils décrivent aussi les différentes manières de parenthéser un mot $m_1m_2\cdots m_n$: pour le mot $abcd$, on ainsi les cinq possibilités $\big((ab)c\big)d$, $\big(a(bc)\big)d$, $a\big((bc)d\big)$ $a\big(b(cd)\big)$ et $(ab)(cd)$. Nous allons compter le nombre a_n d'arbres binaires ayant n sommets terminaux.

Soient A et B des arbres binaires à p et q sommets terminaux. Appelons a le sommet initial de A et b le sommet initial de B. Formons un nouvel arbre $A*B$ en nous donnant son sommet initial S et en accrochant A et B par leur sommet a et b à deux arcs $\widehat{S,a}$ et $\widehat{S,b}$ issus de S : l'arbre binaire $A*B$ possède $p+q$ sommets terminaux. Tout arbre binaire est formé de cette manière, à condition d'introduire parmi les arbres binaires l'arbre-point qui n'a qu'un seul sommet et pas d'arête. On en déduit que le nombre a_n est la somme

des produits $a_p a_q$, où p et q sont quelconques tels que $p{+}q=n$. Il y a un seul un arbre binaire n'ayant que deux sommets terminaux, donc $a_2=1$. Puisque $a_1=1$, on obtient

$$a_n = a_1 a_{n-1} + a_2 a_{n-2} + \cdots + a_{n-1} a_1 \ , \quad \text{pour tout entier } n \geqslant 3.$$

Considérons la série entière $\sum a_n x^n$, où l'on a posé $a_0=0$. Dans la série produit $(\sum a_n x^n)^2$, le coefficient b_n de x^n est précisément la somme ci-dessus, d'après la définition page 535. On a donc

$$b_0 = b_1 = 0 = a_0 \ , \quad b_2 = a_1^2 = 1 = a_2 \quad \text{et} \quad b_n = a_n \ \text{pour } n \geqslant 3.$$

Si l'on admet pour le moment que la série définit bien une fonction $S(x) = \sum_{n=0}^{+\infty} a_n x^n$, alors il vient

$$S(x) - \big[S(x)\big]^2 = \sum_{n=0}^{+\infty} (a_n - b_n) x^n = (a_1 - b_1) x = x\,.$$

Ainsi $S(x)$ est solution de l'équation $Z^2 - Z - x$ et $S(0)=0$, d'où $S(x) = \frac{1}{2} - \frac{1}{2}\sqrt{1-4x}$.

Dans le développement en série de $\sqrt{1-4x}$, le coefficient de x^n est $-\dfrac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots 2n}4^n$ (page 541), donc $a_n = \dfrac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots 2n}2^{2n-1} = \dfrac{1\cdot 3\cdots(2n-3)}{1\cdot 2\cdots n}2^{n-1}$. En multipliant haut et bas par $2\cdot 4\cdots(2n-2) = 2^{n-1}(n-1)!$, il vient la formule simple

$$a_n = \frac{(2n-2)!}{n!(n-1)!} = \frac{1}{2n-1}\binom{2n-1}{n} \ , \quad \text{pour tout } n \geqslant 1.$$

On dit que $S(x)$ est la *fonction génératrice* de la suite (a_n), car le développement en série de $S(x)$ a pour coefficients les a_n.

3. Décomposition de Fourier

Nous allons considérer des fonctions f d'une variable réelle, continues, périodiques de période T et à valeurs éventuellement complexes : on a donc $f(x{+}T) = f(x)$ pour tout $x \in \mathbb{R}$. Soit V l'espace vectoriel constitué par ces fonctions.

Si f et g sont des fonctions appartenant à V, on définit leur produit hermitien comme page 207, en posant

$$f \cdot g = \frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt$$

La norme associée est $\|f\| = \left[\frac{1}{T}\int_0^T \big|f(t)\big|^2\,dt\right]^{1/2}$.

Si par exemple f est un signal temporel, $\|f\|^2$ s'interprète comme la puissance dissipée pendant une période.

Posons $\omega = \dfrac{2\pi}{T}$. Pour tout entier $n \in \mathbb{Z}$, définissons la fonction e_n en posant

$$e_n(x) = e^{n\boldsymbol{i}\omega x} \ , \text{ pour tout } x \in \mathbb{R}.$$

Les fonctions e_n ont pour période T : en effet, $e_n(T)=e^{n\,\boldsymbol{i}\,\omega T}=e^{2\pi\,\boldsymbol{i}}=1$, donc pour tout x, on a $e_n(x+T)=e_n(x)e_n(T)=e_n(x)$. De plus, nous avons montré (page 208) que

$$e_n \cdot e_p = \begin{cases} 0 & \text{si } n \neq p \\ 1 & \text{si } n = p \end{cases}, \quad \text{autrement dit :}$$

les fonctions $e_n(x)=e^{n\,\boldsymbol{i}\,\omega x}$*, où* $n \in \mathbb{Z}$*, forment une famille orthonormée dans* V*.*

Définition

Soit f une fonction appartenant à V. Les *coefficients de Fourier* de f sont les nombres complexes $c_n(f) = f \cdot e_n = \dfrac{1}{T}\displaystyle\int_0^T f(t)\, e^{-n\,\boldsymbol{i}\,\omega t}\, dt$, où $n \in \mathbb{Z}$.

Rappel : Si u est une fonction périodique de période T, l'intégrale $\displaystyle\int_0^T u(t)\, dt$ est égale à l'intégrale de u sur n'importe quel intervalle de longueur T.

3.1 Approximation par des polynômes trigonométriques

Pour tout entier $N > 0$, notons V_N l'espace vectoriel ayant pour base les e_n avec $-N \leqslant n \leqslant N$. Les éléments de V_N sont les combinaisons linéaires à coefficients complexes des fonctions e_n pour $|n| \leqslant N$. Ils s'écrivent de manière unique

$$P(x) = \textstyle\sum_{n=-N}^{N} \lambda_n e_n(x), \quad \text{où les } \lambda_n \text{ sont des nombres complexes.}$$

De telles fonctions $P(x)$ s'appellent des *polynômes trigonométriques*, car pour tout entier relatif n, on a $e_n(x) = \left(e^{\boldsymbol{i}\,\omega x}\right)^n = \left(e_1(x)\right)^n$.

Pour un polynôme trigonométrique $P = \sum_{n=-N}^{N} \lambda_n e_n$*, on a* $\|P\|^2 = \sum_{n=-N}^{N} |\lambda_n|^2$.

En effet, les λ_n sont les coordonnées de P dans la base orthonormée des e_n (proposition page 206).

Soit $f \in V$. D'après le théorème de la projection page 209, le projeté orthogonal de f sur le sous-espace V_N est le polynôme trigonométrique

$$S_N(x) = \textstyle\sum_{n=-N}^{N} c_n(f) e_n(x)\,.$$

On a donc : $\|f - S_N\| \leqslant \|f - P\|$, pour tout $P \in V_N$.

Le polynôme trigonométrique S_N *est la meilleure approximation de* f*, au sens de la norme de* V*, par un polynôme trigonomérique ne faisant intervenir que les fonctions* $e^{n\,\boldsymbol{i}\,\omega x}$ *avec* $|n| \leqslant N$*.*

Au moyen de polynômes trigonométriques, on peut approcher les valeurs de f sur tout l'intervalle $[0,T]$, car on a la propriété suivante.

Etant donné un nombre $\varepsilon > 0$*, il existe un polynôme trigonométrique* P *tel que* $\left|f(x) - P(x)\right| \leqslant \varepsilon$ *quel que soit* $x \in [0,T]$*.*

Nous verrons que ce qui compte vraiment, c'est l'existence de polynômes trigonométriques qui rendent $\|f - P\|$ aussi petit qu'on veut.

Propriétés des coefficients de Fourier

Puisque S_N est le projeté de f sur V_N, $f-S_N$ est orthogonal à V_N, donc à S_N. On a $f=(f-S_N)+S_N$, où f et $f-S_N$ sont orthogonaux, donc par le théorème de Pythagore (page 203), il vient $\|f\|^2=\|f-S_N\|^2+\|S_N\|^2 \geqslant \|S_N\|^2$. Comme $\|S_N\|^2=\sum_{n=-N}^{N}\left|c_n(f)\right|^2$, on en déduit

$$\sum_{n=-N}^{N}\left|c_n(f)\right|^2 \leqslant \|f\|^2 \text{ , pour tout } N.$$

Nous allons être amenés à sommer des nombres indexés par tous les entiers, positifs ou négatifs. Étant donnés des nombres complexes $(a_n)_{n\in\mathbb{Z}}$, nous dirons que la série $\sum a_n$ est convergente si les sommes partielles $\sum_{n=-N}^{N} a_n$ ont une limite quand N tend vers $+\infty$. La somme se note alors $\sum_{n=-\infty}^{+\infty} a_n$. De même, une série $\sum_{n\in\mathbb{Z}} a_n$ est dite absolument convergente si $\sum |a_n|$ est convergente.

Théorème de Bessel. *Pour toutes fonctions f et g continues et périodiques de période T, on a*

i) $\displaystyle\lim_{N\to+\infty} \|f-\sum_{n=-N}^{N} c_n(f)e_n\|=0$

ii) $\displaystyle\frac{1}{T}\int_0^T \left|f(t)\right|^2 dt=\sum_{n=-\infty}^{+\infty}\left|c_n(f)\right|^2$ *(égalité de Bessel)*

iii) $\displaystyle\frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt=\sum_{n=-\infty}^{+\infty} c_n(f)\overline{c_n(g)}$.

D'après (ii), la série $\sum\left|c_n(f)\right|^2$ est convergente, donc son terme général tend vers 0.

Les coefficients de Fourier $c_n(f)$ tendent vers 0 quand n tend vers $+\infty$ ou vers $-\infty$.

Démonstration du théorème. La série $\sum\left|c_n(f)\right|^2$ est à termes positifs et ses sommes partielles $\sum_{n=-N}^{N}\left|c_n(f)\right|^2$ sont toutes majorées par $\|f\|^2$, donc la série converge. Pour tout N, posons $S_N=\sum_{n=-N}^{N} c_n(f)e_n$. Choisissons un polynôme trigonométrique P tel que $\left|f(x)-P(x)\right|\leqslant\varepsilon$ pour tout $x\in[0,T]$. On a alors $\|f-P\|^2=\dfrac{1}{T}\displaystyle\int_0^T\left|f(t)-P(t)\right|^2 dt\leqslant\varepsilon^2$. Puisque P est combinaison linéaire d'un nombre fini de e_k, il y a un indice q tel que P est combinaison des e_k pour $|k|\leqslant q$. Pour tout $N\geqslant q$, S_N-P est alors combinaison des e_k pour $|k|\leqslant N$, autrement dit $S_N-P\in V_N$. D'autre part, $f-S_N$ est orthogonal à V_N et l'on a $(f-S_N)+(S_N-P)=f-P$. D'après le théorème de Pythagore, on en déduit

$$\|f-S_N\|^2+\|S_N-P\|^2=\|f-P\|^2$$

Par suite, on a $\|f-S_N\|\leqslant\|f-P\|\leqslant\varepsilon$ pour tout $N\geqslant q$, ce qui montre (i).
On a aussi $f=(f-S_N)+S_N$ et d'après le théorème de Pythagore, $\|f\|^2=\|f-S_N\|^2+\|S_N\|^2$. Or nous venons de voir que $\|f-S_N\|$ tend vers 0 quand N tend vers $+\infty$, donc $\lim_{N\to+\infty}\|S_N\|^2=\|f\|^2$, ce qui est l'égalité (ii).
Notons S'_N le projeté orthogonal de la fonction g sur V_N. On a $(f-S_N)\cdot S'_N=0=S_N\cdot(g-S'_N)$, car $f-S_N$ et $g-S'_N$ sont orthogonaux à V_N. Par suite $(f-S_N)\cdot(g-S'_N)=(f\cdot g)-(S_N\cdot S'_N)$. Puisque le module du produit scalaire est inférieur ou égal au produit des normes (inégalité de Cauchy-Schwarz, page 204), on en déduit $\left|(f\cdot g)-(S_N\cdot S'_N)\right|\leqslant\|f-S_N\|\,\|g-S'_N\|$. Le membre de droite tend vers 0 quand N tend vers $+\infty$, donc aussi $(f\cdot g)-(S_N\cdot S'_N)$. Puisque S_N et S'_N

sont des polynômes trigonométriques, leur produit scalaire est $S_N \cdot S'_N = \sum_{n=-N}^{N} c_n(f)\overline{c_n(g)}$, d'où l'égalité (iii). ■

Proposition. *Soient f et g des fonctions continues et périodiques de période T.*

i) *Pour tout n, on a $c_n(f+g) = c_n(f) + c_n(g)$ et si λ est un nombre complexe, $c_n(\lambda f) = \lambda c_n(f)$.*

ii) *Si $c_n(f) = c_n(g)$ pour tout $n \in \mathbb{Z}$, alors $f(x) = g(x)$ quel que soit x.*

Démonstration. La première propriété vient de la linéarité de l'intégrale. Supposons $c_n(f) = c_n(g)$ quel que soit $n \in \mathbb{Z}$, donc pour la fonction $u = f - g$, on a $c_n(u) = 0$ quel que soit n. D'après l'égalité de Bessel, $\|u\|^2 = \frac{1}{T}\int_0^T \left|u(t)\right|^2 dt = 0$, donc u est identiquement nulle. ■

Cas d'une fonction à valeurs réelles

Supposons que f est continue, de période $T = \frac{2\pi}{\omega}$ et à valeurs réelles.
On a alors $\overline{f(x)e^{n\boldsymbol{i}\omega x}} = f(x)e^{-n\boldsymbol{i}\omega x}$ et par définition des coefficients de Fourier, il vient $\overline{c_n(f)} = c_{-n}(f)$. On pose pour tout $n \geqslant 0$:

$$a_n(f) = \frac{2}{T}\int_0^T f(t)\cos n\omega t\, dt \quad \text{et} \quad b_n(f) = \frac{2}{T}\int_0^T f(t)\sin n\omega t\, dt.$$

Puisque $e^{n\boldsymbol{i}\omega x} = \cos n\omega x + \boldsymbol{i}\sin n\omega x$, on a alors (en omettant la référence à f)

$$a_n = c_n + c_{-n} \qquad b_n = \boldsymbol{i}(c_n - c_{-n})\ , \text{ donc } a_0 = 2c_0 \text{ et } b_0 = 0$$
$$2c_n = a_n - \boldsymbol{i}\, b_n \qquad 2c_{-n} = a_n + \boldsymbol{i}\, b_n = 2\overline{c_n}$$

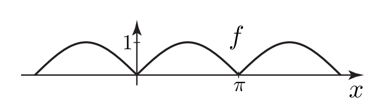
d'où $2\left[|c_n|^2 + |c_{-n}|^2\right] = a_n^2 + b_n^2$. On en déduit les formulations suivantes.

- $\frac{2}{T}\int_0^T \left|f(t)\right|^2 dt = \frac{a_0^2}{2} + \sum_{n=1}^{+\infty}(a_n^2 + b_n^2)$ (*égalité de Bessel*).
- $S_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N}(a_n \cos n\omega x + b_n \sin n\omega x)$.
- $\frac{2}{T}\int_0^T \left|f(t) - S_N(t)\right|^2 dt = \sum_{n=N+1}^{+\infty}(a_n^2 + b_n^2)$ tend vers 0 quand N tend vers $+\infty$.

Parité. On a $b_n(f) = \int_{-T/2}^{T/2} f(t)\sin n\omega t\, dt$: si la fonction f est paire, alors $f(t)\sin n\omega t$ est impaire et l'intégrale précédente est nulle. De même, si f est impaire, $f(t)\cos n\omega t$ est impaire et $a_n(f) = 0$.

Si la fonction f est paire, alors $b_n(f) = 0$ pour tout n.
Si la fonction f est impaire, alors $a_n(f) = 0$ pour tout n.

Exemple. Posons $f(x) = |\sin x|$. La fonction f est continue et puisque $\sin(x+\pi) = -\sin x$, f est périodique de période π : on a ici $T = \pi$, donc $\omega = 2$. La



fonction f étant paire, ses coefficients b_n sont tous nuls. On a

$$a_n = \frac{2}{\pi}\int_0^{\pi} \sin t \cos 2nt\, dt = \frac{1}{\pi}\int_0^{\pi} \big[\sin(2n+1)t + \sin(2n-1)t\big]\, dt = -\frac{4}{\pi}\,\frac{1}{4n^2-1}$$

En particulier, $a_0 = \frac{4}{\pi}$ et l'égalité de Bessel s'écrit

$$\frac{8}{\pi^2} + \frac{16}{\pi^2}\sum_{n=1}^{+\infty} \frac{1}{(4n^2-1)^2} = \frac{2}{\pi}\int_0^{\pi} (\sin t)^2\, dt$$

On a

$$\int_0^{\pi} (\sin t)^2\, dt = \frac{1}{2}\int_0^{\pi} (1-\cos 2t)\, dt = \frac{\pi}{2} - \frac{1}{4}\big[\sin 2t\big]_0^{\pi} = \frac{\pi}{2}$$

d'où $\frac{8}{\pi^2} + \frac{16}{\pi^2}\sum_{n=1}^{+\infty} \frac{1}{(4n^2-1)^2} = 1$ et finalement $\sum_{n=1}^{+\infty} \frac{1}{(4n^2-1)^2} = \frac{\pi^2}{16} - \frac{1}{2}$.

Approximations en moyenne

1) D'après le théorème de Bessel, les polynômes trigonométriques $S_N(x)=\sum_{n=-N}^{N} c_n(f)e^{n\boldsymbol{i}\omega x}$ ont la propriété : $\int_0^T \big|f(t)-S_N(t)\big|^2\, dt$ tend vers 0 quand N tend vers $+\infty$.

On dit que les fonctions S_N *tendent vers* f *en moyenne quadratique.*

2) Montrons que $\int_0^T \big|f(t)-S_N(t)\big|\, dt$ tend aussi vers 0 quand N tend vers $+\infty$.

On dit que les fonctions S_N *tendent en moyenne vers* f.

Pour toutes fonctions f et g continues sur $[0, T]$, on a l'inégalité de Cauchy-Schwarz (page 207)

$$\left|\int_0^T f(t)\overline{g(t)}\, dt\right| \leqslant \left(\int_0^T |f(t)|^2\, dt\right)^{1/2} \left(\int_0^T |g(t)|^2\, dt\right)^{1/2}$$

Appliquons cette inégalité à la fonction $\big|f(x)\big|$ en prenant g constante de valeur 1 : il vient

$$0 \leqslant \int_0^T \big|f(t)\big|\, dt \leqslant \left(\int_0^T |f(t)|^2\, dt\right)^{1/2} \left(\int_0^T dt\right)^{1/2} = \sqrt{T}\left(\int_0^T |f(t)|^2\, dt\right)^{1/2}$$

Remplaçons f par $f-S_N$: d'après le théorème de Bessel, $\int_0^T \big|f(t)-S_N(t)\big|^2\, dt$ tend vers 0 quand N tend vers $+\infty$. L'intégrale $\int_0^T \big|f(t)-S_N(t)\big|\, dt$ est positive et majorée par une quantité qui tend vers 0 quand N tend vers $+\infty$, donc elle tend vers 0 quand N tend vers $+\infty$.

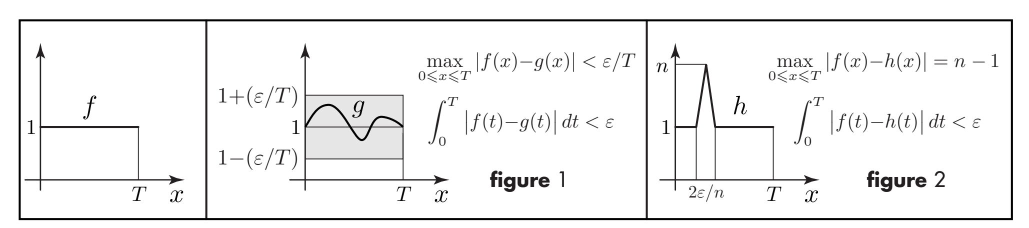
Dans une approximation en moyenne du type $\int_0^T \big|f(t)-g(t)\big|\, dt < \varepsilon$, ce qui compte, c'est que l'aire comprise entre le graphe de f et celui de g soit petite en valeur absolue : les fonctions f et g peuvent prendre des valeurs très différentes, mais seulement sur des intervalles très petits.

Une approximation en moyenne n'est pas nécessairement une approximation en valeurs.

Illustrons les deux types d'approximation pour la fonction f constante de valeur 1 sur $[0,T]$.

► La figure 1 montre une approximation en valeurs : le graphe de g est dans la bande $1-(\varepsilon/T) < g(x) < 1+(\varepsilon/T)$, donc $\left|f(x)-g(x)\right| < \varepsilon/T$ quel que soit x. Par suite, on a $\int_0^T \left|f(t)-g(t)\right| dt \leqslant \int_0^T \varepsilon/T\, dt = \varepsilon$. Sur un segment, une petite approximation en valeurs est aussi une petite approximation en moyenne.

► Sur la figure 2, on a pris $h(x) = f(x)$ sauf sur un petit segment de longueur $2\varepsilon/n$ et le maximum de h est choisi égal à n, un entier strictement positif quelconque : la fonction h peut prendre de grandes valeurs, mais l'intervalle où cela se produit a une longueur petite. En moyenne, l'approximation $\int_0^T \left|f(t)-h(t)\right| dt$ est l'aire d'un triangle de base $2\varepsilon/n$ et de hauteur $n-1$, donc $\int_0^T \left|f(t)-h(t)\right| dt = (1/2)(2\varepsilon/n)(n-1) < \varepsilon$.



3.2 La décomposition de Fourier

élargissons le cadre précédent pour tenir compte de fonctions ayant des discontinuités ou qui ne sont pas définies en certains points. On doit se limiter à des fonctions u telles que l'intégrale généralisée $\int_0^T \left|u(t)\right|^2 dt$ existe.

Définition

Si l'intégrale généralisée $\int_0^T \left|u(t)\right|^2 dt$ existe, on dit que la fonction u est *de carré intégrable sur* $[0,T]$. L'ensemble de ces fonctions forme un espace vectoriel noté $\mathbb{L}^2\big([0,T]\big)$.

Exemple. Posons $u(x) = \dfrac{\cos x}{x^\alpha}$ pour $0 < x < 2\pi$. On a $\left|u(x)\right|^2 \leqslant \dfrac{1}{x^{2\alpha}}$ et si $2\alpha < 1$, l'intégrale généralisée $\int_0^{2\pi} \dfrac{1}{t^{2\alpha}}\, dt$ existe (page 327). On en déduit que si $\alpha < 1/2$, la fonction u est de carré intégrable sur $[0,2\pi]$.

Si l'on part d'une fonction u continue sur $[0,T]$ et qu'on la modifie en un nombre fini de points, on obtient une fonction v telle que $|u-v|$ et $|u-v|^2$ ont une intégrale nulle. Pour des fonctions non continues, la propriété $\int_0^T \left|u(t)-v(t)\right|^2 dt = 0$ n'assure donc pas que les fonctions u et v sont égales, mais seulement que, sur $[0,T]$, on a $u(x) = v(x)$ sauf en des points formant un ensemble négligeable du point de vue de l'intégrale, comme par

exemple un ensemble fini ou l'ensemble des points d'une suite. Ainsi, dans un espace contenant des fonctions discontinues, l'intégrale $\int_0^T |u(t)|^2\,dt$ ne constitue plus une norme.

Convention : Dans l'espace $\mathbb{L}^2([0,T])$, convenons de considérer comme identiques des fonctions u et v telles que $\int_0^T |u(t)-v(t)|^2\,dt = 0$.

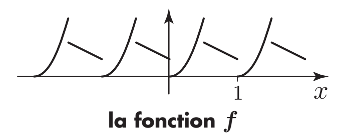
Cette égalité dans $\mathbb{L}^2([0,T])$ s'écrit simplement $u=v$, bien qu'on ait seulement $u(x)=v(x)$ *presque partout*. Mais si l'on veut exprimer que les fonctions u et v prennent les mêmes valeurs en tout point de $[0,T]$, c'est-à-dire sont égales au sens habituel, on écrira : $u(x) = v(x)$ quel que soit $x \in [0,T]$.

Notation : L'ensemble des fonctions périodiques de période T et de carré intégrable sur $[0,T]$ se note $\mathbb{L}_p^2([0,T])$. Pour f et g dans $\mathbb{L}_p^2([0,T])$,

- le produit scalaire $f \cdot g = \frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt$ est bien défini
- et $\|f\| = (f \cdot f)^{1/2}$ est une norme sur $\mathbb{L}_p^2([0,T])$.

Exemple. Considérons la fonction u définie sur $[0,1[$ par

$$u(x) = \begin{cases} 3(e^{x^2}-1) & \text{si } 0 \leqslant x < 1/2 \\ (3-2x)/4 & \text{si } 1/2 < x < 1 \end{cases}.$$

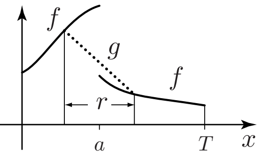


la fonction f

Bien que u ne soit pas continue en $1/2$, elle l'est sur chacun des intervalles $]0,1/2[$ et $]1/2,1[$, donc $u \in \mathbb{L}^2([0,1])$. Étendons u par périodicité en une fonction f de période 1 : alors f appartient à l'espace $\mathbb{L}_p^2([0,1])$.

Théorème. *Les propriétés énoncées dans le théorème de Bessel sont encore vraies pour les fonctions f périodiques et de carré intégrable sur $[0,T]$.*

Ce qui compte en effet dans la démonstration, c'est la possibilité pour tout $\varepsilon > 0$ donné, de trouver un polynôme trigonométrique P tel que $\|f-P\| < \varepsilon$, c'est-à-dire d'approcher f en moyenne quadratique par des polynômes trigonométriques.



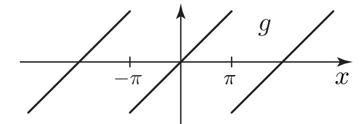
Expliquons comment réaliser cette approximation lorsque f appartient à $\mathbb{L}^2([0,1])$. Prenons par exemple la fonction f dont le graphe a l'allure ci-contre (dessinée en trait plein), avec une discontinuité en a. Modifions f sur un petit intervalle de longueur r autour du point de discontinuité, en considérant la fonction continue g qui emprunte le segment de droite en pointillés : comme f et g ne diffèrent que sur un intervalle de longueur r, l'intégrale $\frac{1}{T}\int_0^T |f(t)-g(t)|^2\,dt$ peut être rendue inférieure à ε^2 à condition de prendre r assez petit : on a alors $\|f-g\| < \varepsilon$. Puisque g est continue, il y a un polynôme trigonométrique P tel que $|g(x)-P(x)| < \varepsilon$ pour tout $x \in [0,T]$, ce qui entraîne $\|g-P\| = \left[\frac{1}{T}\int_0^T |g(t)-P(t)|^2\,dt\right]^{1/2} < \varepsilon$. Comme $f-P = (f-g) + (g-P)$, on a l'inégalité triangulaire $\|f-P\| \leqslant \|f-g\| + \|g-P\|$, d'où

$\|f-P\| \leqslant 2\varepsilon$. Pour des valeurs ε petites, les polynômes P ainsi construits approchent f en moyenne quadratique.

Proposition. *Si f et g appartiennent à $\mathbb{L}_p^2\big([0,T]\big)$ et si $c_n(f) = c_n(g)$ quel que soit n, alors $f = g$ dans $\mathbb{L}_p^2\big([0,T]\big)$.*

En effet, si $c_n(f) = c_n(g)$ quel que soit n, alors $c_n(f-g) = c_n(f) - c_n(g) = 0$, donc $\|f-g\| = 0$ d'après l'égalité de Bessel. Par suite $f = g$ dans $\mathbb{L}_p^2\big([0,T]\big)$.

Exemple. Soit g la fonction périodique de période 2π définie en posant $g(x) = x$ pour $-\pi < x \leqslant \pi$. On a ici $\omega = 2\pi/T = 1$ et comme g est impaire, ses coefficients a_n sont tous nuls. On a $b_n = \dfrac{2}{2\pi}\displaystyle\int_{-\pi}^{\pi} t \sin nt \, dt$ et



$$\int_{-\pi}^{\pi} t \sin nt \, dt = \frac{-1}{n}\big[t\cos nt\big]_{-\pi}^{\pi} + \frac{1}{n}\int_{-\pi}^{\pi} \cos nt \, dt$$

$$= \frac{-1}{n}\big[2(-1)^n\pi\big] = \frac{(-1)^{n+1}}{n}2\pi \text{ car la dernière intégrale est nulle,}$$

donc $b_n = 2\dfrac{(-1)^{n+1}}{n}$. Explicitons l'égalité de Bessel $\dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} \big|g(t)\big|^2 \, dt = \sum_{n=1}^{+\infty} b_n^2$. On a $\dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} t^2 \, dt = \dfrac{1}{\pi}\dfrac{2\pi^3}{3} = \dfrac{2\pi^2}{3}$ et $b_n^2 = \dfrac{4}{n^2}$, d'où l'égalité $\sum_{n=1}^{+\infty} \dfrac{1}{n^2} = \dfrac{\pi^2}{6}$.

La décomposition de Fourier

Définition

Une famille $(\alpha_n)_{n\in\mathbb{Z}}$ de nombre complexes est dite *de carré sommable* si la série $\sum |\alpha_n|^2$ est convergente. Ces suites constituent un espace vectoriel noté ℓ^2.

Si $f \in \mathbb{L}_p^2\big([0,T]\big)$, la série $\sum \big|c_n(f)\big|^2$ est convergente, donc $\big(c_n(f)\big)_{n\in\mathbb{Z}}$ est un élément de ℓ^2.

Définition

La *décomposition de Fourier* est l'application $\mathscr{F} : \mathbb{L}_p^2\big([0,T]\big) \to \ell^2$ qui à f associe $\big(c_n(f)\big)_{n\in\mathbb{Z}}$.

Dans ce cadre plus général, on a le résultat suivant.

Théorème. *La décomposition de Fourier $\mathscr{F} : \mathbb{L}_p^2\big([0,T]\big) \to \ell^2$ est une bijection linéaire.*

Justification. Puisque chaque coefficient de Fourier c_n dépend linéairement de la fonction, l'application $\mathscr{F}$ est linéaire. Si $\mathscr{F}(f) = \mathscr{F}(g)$, alors $c_n(f) = c_n(g)$ que que soit n et d'après la proposition précédente, on a $f = g$ dans $\mathbb{L}_p^2[0,T]$: par la décomposition de Fourier, tout élément de ℓ^2 a donc au plus un antécédent.
Admettons que si $(a_n)_{n\in\mathbb{Z}}$ est un élément donné de ℓ^2, il y a une fonction $f \in \mathbb{L}_p^2\big([0,T]\big)$ telle

que $\|f - P_N\|$ tend vers 0, où les P_N sont les polynômes trigonométriques $\sum_{n=-N}^{N} a_n e^{n\boldsymbol{i}\omega x}$. Soit n un entier tel que $|n| \leqslant N$. On a $P_N \cdot e_n = a_n$, car les e_k sont deux à deux orthogonaux et de norme 1. Par suite $\left|c_n(f) - a_n\right| = \left|f \cdot e_n - P_N \cdot e_n\right| = \left|(f - P_N) \cdot e_n\right| \leqslant \|f - P_N\| \|e_n\| = \|f - P_N\|$. Quand N tend vers $+\infty$, le membre de droite tend vers 0, donc $\left|c_n(f) - a_n\right| = 0$ et $c_n(f) = a_n$. On a ainsi $\mathscr{F}(f) = (a_n)_{n \in \mathbb{Z}}$: tout élément de ℓ^2 est donc la décomposition de Fourier d'une fonction de $\mathbb{L}_p^2\big([0,T]\big)$. ■

Par décomposition de Fourier, un signal périodique f de carré intégrable se code en une suite $\big(c_n(f)\big)_{n \in \mathbb{Z}}$ de carré sommable.
Le code caractérise parfaitement le signal et de plus, toute suite $(\alpha_n)_{n \in \mathbb{Z}}$ de carré sommable est le code d'un unique signal périodique.

3.3 Propriétés de la décomposition de Fourier

Notation : Pour toute suite $(\alpha_n)_{n \in \mathbb{Z}}$ de carré sommable, on note $\sum_{n=-\infty}^{+\infty} \alpha_n e_n$ la fonction de $\mathbb{L}_p^2\big([0,T]\big)$ ayant pour coefficients de Fourier les α_n, c'est-à-dire la fonction f telle que $\mathscr{F}(f) = (\alpha_n)_{n \in \mathbb{Z}}$.
D'après le théorème précédent, on a donc

$$f = \sum_{n=-\infty}^{+\infty} c_n(f) e_n, \quad \text{pour tout } f \in \mathbb{L}_p^2\big([0,T]\big).$$

Cette notation fait apparaître f comme la somme, dans l'espace $\mathbb{L}_p^2\big([0,T]\big)$, de la série $\sum c_n(f) e_n$.

Définitions
La série $\sum c_n(f) e_n$ s'appelle la *série de Fourier* de f.

Les fonctions $c_n(f) e_n(x) = c_n(f) e^{n\boldsymbol{i}\frac{2\pi}{T}x}$, qu'on appelle *les harmoniques de f*, constituent la *décomposition spectrale* de f.

- Le coefficient $c_0 = (1/T) \int_0^T f(t)\, dt$ est la moyenne de f sur une période.
- L'harmonique $c_n e^{\boldsymbol{i}\, 2\pi n x / T}$ est de période T/n, donc de fréquence n fois la fréquence $1/T$ de f. Son *amplitude* est $|c_n|$.
- Dans $\mathbb{L}_p^2\big([0,T]\big)$, l'égalité $f = \sum_{n=-\infty}^{+\infty} c_n e_n$ signifie qu'en tenant compte d'un nombre suffisant d'harmoniques, on peut approcher le signal f par une somme finie $S_N = \sum_{n=-N}^{N} c_n e_n$ de manière que sur chaque période, $f - S_N$ soit d'énergie aussi petite qu'on veut.
- Si la fonction f est à valeurs réelles, elle est égale dans $\mathbb{L}_p^2\big([0,T]\big)$ à

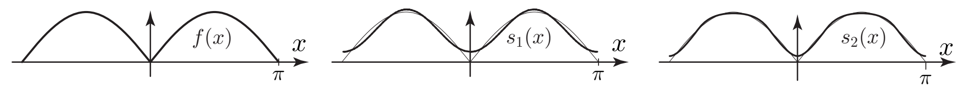
$$\frac{a_0}{2} + \sum_{n=1}^{+\infty} \big(a_n \cos n\omega x + b_n \sin n\omega x\big).$$

  La fonction $a_1 \cos \omega x + b_1 \sin \omega x$, de période T, est *l'harmonique principal* de f.

Exemple 1 : approximation d'un courant redressé. La fonction $f(x)=|\sin x|$ a pour série de Fourier $\frac{2}{\pi}-\frac{4}{\pi}\sum_{n=1}^{+\infty}\frac{\cos 2nx}{4n^2-1}$ (exemple page 551). Sur les figures ci-dessous, on a représenté le graphe de f et celui des deux premières approximations

$$s_1(x)=\frac{2}{\pi}-\frac{4}{\pi}\frac{\cos 2x}{3} \qquad \text{et} \qquad s_2(x)=\frac{2}{\pi}-\frac{4}{\pi}\left[\frac{\cos 2x}{3}+\frac{\cos 4x}{15}\right].$$

Remarquons que ces approximations sont très bonnes, bien que, en valeurs, elles le soient un peu moins au voisinage des points $k\pi$ où le graphe présente deux demi-tangentes de pentes différentes.



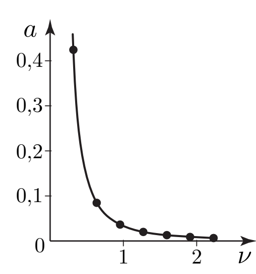
Quand on redresse un courant alternatif de fréquence ν, on obtient pour la tension une fonction du temps de la forme $V(t)=V_{\max}\left|\sin(2\pi\nu t+\varphi)\right|$. Les premiers termes de la série de Fourier permettent ainsi d'approcher V par des polynômes trigonométriques de fréquences $\nu, 2\nu, 3\nu, \ldots$

Représentation spectrale. Commençons par centrer le signal $f(x)=|\sin x|$ précédent autour de sa moyenne en considérant $f_c(x)=f(x)-\frac{2}{\pi}$ qui est de moyenne nulle. Puisque f_c est de période π, sa fréquence est $\nu_0=1/\pi$.

Considérons un axe des fréquences et pour chacune des abscisses $\nu_0, 2\nu_0, \ldots, k\nu_0, \ldots$, plaçons en ordonnée l'amplitude $|a_k|$. On a ainsi $|a_1|=\frac{4}{3\pi}$, $|a_2|=\frac{4}{15\pi}$ et $|a_n|=\frac{4}{(4n^2-1)\pi}$.

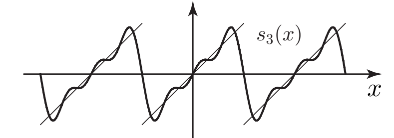
Le graphique obtenu est une *décomposition en fréquences* du signal f_c (figure ci-contre).

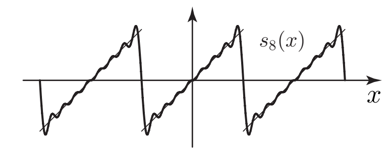
Pour l'abscisse $\nu=k\nu_0$, il vient $k=\pi\nu$: les points sont donc sur la courbe d'équation $a=\frac{4}{\pi}\frac{1}{4\pi^2\nu^2-1}$, où a est l'amplitude.



Exemple 2. Reprenons la fonction périodique de période 2π définie en posant $g(x)=x$ pour tout $x\in]-\pi,\pi]$. D'après les calculs faits dans l'exemple page 555, sa série de Fourier est $2\sum_{n=1}^{+\infty}(-1)^{n+1}\frac{\sin nx}{n}$. Voici les graphes de g et des approximations en moyenne

$$s_3(x)=2\sum_{n=1}^{3}(-1)^{n+1}\frac{\sin nx}{n} \qquad \text{et} \qquad s_8(x)=2\sum_{n=1}^{8}(-1)^{n+1}\frac{\sin nx}{n}.$$



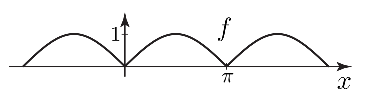


Puisque la fonction est discontinue aux points $\pi+2k\pi$, l'approximation en valeurs y est nécessairement mauvaise. Remarquons aussi que sur de petits intervalles comme $]\pi-h,\pi[$ et $]\pi,\pi+h[$, l'approximation se dégrade : c'est un phénomène courant.

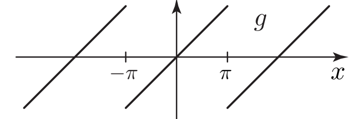
Dérivation et décomposition de Fourier

On rencontre souvent des fonctions qui, sur $[0,T]$, sont dérivables sauf peut-être en un nombre fini de points. Si de plus la dérivée est continue et bornée sur les intervalles ouverts où elle est définie, on dit que f est *continûment dérivable par morceaux*. Puisque nous supposons que la dérivée est bornée, f' est de carré intégrable sur $[0,T]$. Voici des exemples de fonctions continûment dérivables par morceaux.

1) Pour la fonction $f(x) = |\sin x|$, on a $f'(x) = \cos x$ si $x \in]0,\pi[$, mais aux points $0, \pm\pi, \ldots, \pm k\pi \ldots$, f n'est pas dérivable : il y a deux demi-tangentes de pentes 1 et -1. La fonction f est continue, périodique de période π et continûment dérivable par morceaux.



2) La fonction périodique de période 2π définie en posant $g(x) = x$ pour tout $x \in]-\pi,\pi]$ n'est pas continue en π, non plus qu'aux points $\pi + 2k\pi$, où $k \in \mathbb{Z}$. Néanmoins, la fonction g est continûment dérivable par morceaux, car pour $-\pi < x < \pi$, la dérivée $g'(x) = 1$ est continue.



Proposition. *Si f est une fonction périodique de période T, continue et continûment dérivable par morceaux, alors $c_n(f') = \boldsymbol{i}\, n\omega\, c_n(f)$.*

Démonstration. Supposons par exemple que f a une dérivée continue sur $]0,a[$ et sur $]a,T[$, où a est un nombre tel que $0 < a < T$. En intégrant par parties, il vient

$$\int_0^a f'(t)e^{-n\boldsymbol{i}\omega t}\,dt = \left[f(x)e^{-n\boldsymbol{i}\omega x}\right]_0^a + n\boldsymbol{i}\omega \int_0^a f(t)e^{-n\boldsymbol{i}\omega t}\,dt$$

$$\int_a^T f'(t)e^{-n\boldsymbol{i}\omega t}\,dt = \left[f(x)e^{-n\boldsymbol{i}\omega x}\right]_a^T + n\boldsymbol{i}\omega \int_a^T f(t)e^{-n\boldsymbol{i}\omega t}\,dt\,.$$

Ajoutons ces deux expressions. La somme des crochets est $\left[f(x)e^{-n\boldsymbol{i}\omega x}\right]_0^T$, car f est continue en a, et puisque $f(x)e^{-n\boldsymbol{i}\omega x}$ a pour période T, on a $\left[f(x)e^{-n\boldsymbol{i}\omega x}\right]_0^T = 0$. Il vient ainsi l'égalité $\int_0^T f'(t)e^{-n\boldsymbol{i}\omega t}\,dt = n\boldsymbol{i}\omega \int_0^T f(t)e^{-n\boldsymbol{i}\omega t}\,dt$, d'où $c_n(f') = n\boldsymbol{i}\omega c_n(f)$. ■

Conséquences. *Soit f une fonction périodique.*

- *Si f est continue et continûment dérivable par morceaux, ses coefficients a_n, b_n et c_n sont négligeables devant $1/n$.*
- *Si f possède une dérivée $f^{(k)}$ continue sur $\mathbb{R}$, ses coefficients a_n, b_n et c_n sont négligeables devant $1/n^k$.*

Plus une fonction périodique est dérivable, plus vite ses coefficients de Fourier tendent vers 0.

Démonstration. Si f est continue et continûment dérivable par morceaux, alors d'après la proposition, on a $\left|c_n(f')\right| = \left|\boldsymbol{i}\, n\omega\, c_n(f)\right| = \omega n \left|c_n(f)\right|$. Puisque $c_n(f')$ tend vers 0 quand $|n|$ tend vers $+\infty$, il vient $\lim\limits_{|n|\to+\infty} n\left|c_n(f)\right| = 0$: cela montre que $c_n(f)$ est négligeable devant

$1/n$. Il en va de même de a_n et b_n dont la valeur absolue est majorée par $2|c_n|$.
Supposons que f a une dérivée k-ième continue ($k \geqslant 2$). On a alors $c_n(f'') = \boldsymbol{i}\, n\omega c_n(f') = (\boldsymbol{i}n\omega)^2 c_n(f)$ et de proche en proche, il vient $c_n(f^{(k)}) = (\boldsymbol{i}n\omega)^k c_n(f)$. On sait que $\omega^k n^k \left|c_n(f)\right| = \left|c_n(f^{(k)})\right|$ tend vers 0 quand $|n|$ tend vers $+\infty$, par conséquent $\left|c_n(f)\right| = o(1/n^k)$. ■

Exemple. Pour la fonction $f(x) = \left|\sin x\right|$ qui est continue et continûment dérivable par morceaux, la série de Fourier est $\frac{2}{\pi} - \frac{4}{\pi}\sum_{n=1}^{+\infty} \frac{\cos 2nx}{4n^2-1}$ (exemple 1 page 557). On a $a_n = -\frac{4}{\pi}\frac{1}{4n^2-1}$ et quand n tend vers l'infini, $|a_n| \sim \frac{1}{\pi n^2}$ est négligeable devant $1/n$.

Produit de convolution

Définition

Si f et g sont des fonctions appartenant à $\mathbb{L}_p^2\big([0,T]\big)$, leur *produit de convolution* $f * g$ est la fonction définie par $(f*g)(x) = \dfrac{1}{T}\displaystyle\int_0^T f(t)g(x-t)\,dt$, pour tout $x \in \mathbb{R}$.

C'est encore l'inégalité de Cauchy-Schwarz qui assure que le produit de convolution est bien défini. Comparer cette définition avec celle donnée page 409 pour des fonctions non périodiques.

Propriétés

i) *Le produit de convolution est symétrique et bilinéaire :* $f*g = g*f$, $f*(g+h) = (f*g)+(f*h)$ *et* $f*(\lambda g) = \lambda(f*g)$.

ii) *On a* $(f*g)*h = f*(g*h)$ *(associativité du produit de convolution).*

iii) $f * e_n = c_n(f)e_n$*, pour tout* $n \in \mathbb{Z}$.

iv) *Si* $f*g$ *est de carré intégrable sur* $[0,T]$*, alors* $\mathscr{F}(f*g) = \big[c_n(f)c_n(g)\big]_{n\in\mathbb{Z}}$.

Démonstration. La bilinéarité résulte des propriétés de l'intégrale et la symétrie s'obtient en faisant le changement de variable $u = x-t$. On a $(f*g)(t) = \dfrac{1}{T}\displaystyle\int_0^T f(s)g(t-s)\,ds$, d'où

$$\begin{aligned}
\big[(f*g)*h\big](x) &= \frac{1}{T^2}\int_0^T \left[\int_0^T f(s)g(t-s)\,ds\right] h(x-t)\,dt \\
&= \frac{1}{T^2}\int_0^T f(s)\left[\int_0^T g(t-s)h(x-t)\,dt\right] ds \\
&= \frac{1}{T^2}\int_0^T f(s)\left[\int_{-s}^{T-s} g(u)h(x-s-u)\,du\right] ds \quad \text{en posant } t = s+u \\
&= \frac{1}{T^2}\int_0^T f(s)\left[\int_0^T g(u)h(x-s-u)\,du\right] ds\,,
\end{aligned}$$

car l'intégrale de 0 à T d'une fonction périodique de période T s'obtient en intégrant sur n'importe quel segment de longueur T. Puisque $(g*h)(x-s) = \dfrac{1}{T}\displaystyle\int_0^T g(u)h(x-s-u)\,du$, la dernière expression est $\dfrac{1}{T}\displaystyle\int_0^T f(s)\big[(g*h)(x-s)\big]\,ds = \big[f*(g*h)\big](x)$, d'où l'associativité du

produit de convolution. Pour montrer (iii), écrivons

$$(f * e_n)(x) = \frac{1}{T}\int_0^T f(t)e^{n\,\boldsymbol{i}\,\omega(x-t)}\,dt = \frac{1}{T}\int_0^T e^{n\,\boldsymbol{i}\,\omega x} f(t)e^{-n\,\boldsymbol{i}\,\omega t}\,dt$$
$$= \left[\frac{1}{T}\int_0^T f(t)e^{-n\,\boldsymbol{i}\,\omega t}\,dt\right] e^{n\,\boldsymbol{i}\,\omega x} = c_n(f)e^{n\,\boldsymbol{i}\,\omega x}\,.$$

Supposons $f * g \in \mathbb{L}^2_p\big([0,T]\big)$. D'après (iii), on a alors

$$\big[c_n(f*g)\big]e_n = (f*g)*e_n = f*(g*e_n) = f*\big[c_n(g)e_n\big] = c_n(g)\,(f*e_n) = c_n(g)c_n(f)\,e_n,$$

en utilisant linéarité et associativité. Il s'ensuit $c_n(f*g) = c_n(f)c_n(g)$ pour tout n. ■

Utilisation de la décomposition de Fourier

La décomposition de Fourier s'utilise comme un changement de référentiel, car de nombreux calculs usuels sur les fonctions périodiques ont leur traduction au moyen des coefficients de Fourier.

a) Si f est dérivable, dériver f se traduit en multipliant chaque $c_n(f)$ par $n\,\boldsymbol{i}\,\omega$.

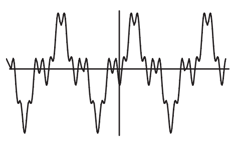
b) Si l'on décale un signal dans le temps d'une quantité a, cela se traduit, pour tout n, par un déphasage de $-n\omega a$ de son harmonique d'indice n :

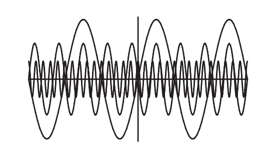
si $f_a(x) = f(x-a)$, on a en effet $c_n(f_a) = e^{-n\,\boldsymbol{i}\,\omega a}c_n(f)$.

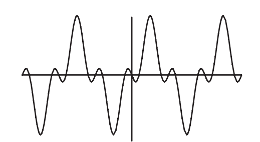
c) Convoler f avec g se traduit en multipliant $c_n(f)$ par $c_n(g)$, à condition que les produits $c_n(f)c_n(g)$ forment une suite de carré sommable.

d) Calculer l'intégrale $\frac{1}{T}\int_0^T \big|f(t)\big|^2\,dt$, ou l'intégrale $\frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt$, revient à calculer la somme $\sum_{n=-\infty}^{+\infty}\big|c_n(f)\big|^2$, ou la somme $\sum_{n=-\infty}^{+\infty} c_n(f)\overline{c_n(g)}$.

Décomposer un signal périodique en série de Fourier permet de le filtrer, c'est-à-dire d'en éliminer les harmoniques de fréquences trop hautes ou trop basses. Les figures ci-dessous montrent un filtrage où l'on élimine l'harmonique de plus haute fréquence.


signal à traiter


les harmoniques du signal


signal filtré

3.4 Convergence ponctuelle d'une série de Fourier

La convergence de la série de Fourier au sens de la norme dans l'espace V, ou dans l'espace $\mathbb{L}^2_p\big([0,T]\big)$, n'implique pas que, pour un nombre réel x donné, la série numérique des $c_n(f)e^{n\,\boldsymbol{i}\,\omega x}$ est convergente. Voici des conditions suffisantes pour que cela soit vrai.

Proposition. *Soit f une fonction continue et périodique de période T. Si la série $\sum\big|c_n(f)\big|$ est convergente, alors pour tout $x \in \mathbb{R}$, on a l'égalité $f(x) = \sum_{n=-\infty}^{+\infty} c_n(f)e^{n\,\boldsymbol{i}\,\omega x}$.*

Démonstration. Pour tout $x \in \mathbb{R}$, on a $\left|c_n(f)e^{n\,\boldsymbol{i}\,\omega x}\right| = \left|c_n(f)\right|$, donc la série $\sum c_n(f)e^{n\,\boldsymbol{i}\,\omega x}$ est absolument convergente. Posons

$$g(x) = \sum_{n=-\infty}^{+\infty} c_n(f)e^{n\,\boldsymbol{i}\,\omega x} = S_N(x) + \sum_{|n|\geqslant N+1} c_n(f)e^{n\,\boldsymbol{i}\,\omega x}, \quad \text{où} \quad S_N(x) = \sum_{n=-N}^{N} c_n(f)e^{n\,\boldsymbol{i}\,\omega x}.$$

La fonction g est périodique de période T. Puisque $\left|c_n(f)e^{n\,\boldsymbol{i}\,\omega x}\right| = \left|c_n(f)\right|$, on a $\left|g(x)-S_N(x)\right| \leqslant \sum_{|n|\geqslant N+1}\left|c_n(f)\right|$ pour tout $x \in \mathbb{R}$. Par hypothèse, la suite $K_N = \sum_{|n|\geqslant N+1}\left|c_n(f)\right|$ tend vers 0, d'où

$$(*) \qquad \text{quel que soit } x \in \mathbb{R}, \quad \left|g(x) - S_N(x)\right| \leqslant K_N, \quad \text{avec} \lim_{N\to+\infty} K_N = 0.$$

Montrons que g est continue en raisonnant comme pour les séries entières, page 536. Étant donné $\varepsilon > 0$, choisissons N pour que $K_N < \varepsilon$. Si x et y sont des nombres réels, on a $\left|e^{n\,\boldsymbol{i}\,\omega x} - e^{n\,\boldsymbol{i}\,\omega y}\right| \leqslant 2\left|\sin\dfrac{x-y}{2}\right|$ (proposition page 39), donc on peut rendre l'écart $\left|S_N(x)-S_N(y)\right|$ moindre que ε à condition de choisir x et y assez proches. D'après l'inégalité $(*)$, l'écart $\left|g(x)-g(y)\right|$ est alors moindre que $\left|g(x)-S_N(x)\right| + \left|S_N(x)(x)-S_N(y)\right| + \left|S_N(y)-g(y)\right| \leqslant 3\varepsilon$. Pour un entier k donné et $N \geqslant |k|$, on a $S_N \cdot e_k = c_k(f)$ par définition de S_N et $g \cdot e_k = c_k(g)$. Il vient $\left|c_k(g) - c_k(f)\right| = \left|(g - S_N)\cdot e_k\right| \leqslant \|g-S_N\|\,\|e_k\| = \|g-S_N\|$. D'après $(*)$, les fonctions S_N approchent g par valeurs sur $\mathbb{R}$, donc la norme $\|g-S_N\|$ tend vers 0 quand N tend vers $+\infty$. Il s'ensuit $c_k(f) = c_k(g)$: les fonctions f et g ont mêmes coefficients de Fourier. Comme f et g sont continues, la proposition page 551 affirme que $f(x) = g(x)$ quel que soit x. ■

Théorème. *Si f est périodique de période T, continue et continûment dérivable par morceaux, alors $f(x) = \sum_{n=-\infty}^{+\infty} c_n(f)e^{n\,\boldsymbol{i}\,\omega x}$ quel que soit $x \in \mathbb{R}$.*

Démonstration. Les coefficients de Fourier de f' sont $c_n(f') = \boldsymbol{i}\,n\omega c_n(f)$. D'après le théorème de Bessel appliqué à la dérivée f', la série $\sum\left|c_n(f')\right|^2$ est convergente, donc aussi la série $\sum\left|nc_n(f)\right|^2$. Mais si u et v sont des nombres réels, on a $2|uv| \leqslant u^2 + v^2$, car $(u \pm v)^2 \geqslant 0$. Appliquons cette inégalité à $u = \left|nc_n(f)\right|$ et $v = \dfrac{1}{n}$, pour $n \neq 0$. Il vient $2\left|c_n(f)\right| = 2\left|nc_n(f)\right|\dfrac{1}{n} \leqslant \left|nc_n(f)\right|^2 + \dfrac{1}{n^2}$. La série de terme général $1/n^2$ est une série de Riemann convergente et la somme de deux séries convergentes est convergente, donc, par le théorème de comparaison page 528, la série $\sum\left|c_n(f)\right|$ est convergente. D'après la proposition précédente, quel que soit le réel x, $f(x)$ est donc la somme de la série de terme général $c_n(f)e^{n\,\boldsymbol{i}\,\omega x}$. ■

Exemple. Pour la fonction $|\sin x|$, on a ainsi l'égalité $|\sin x| = \dfrac{2}{\pi} - \dfrac{4}{\pi}\sum_{n=1}^{+\infty}\dfrac{\cos 2nx}{4n^2-1}$ quel que soit $x \in \mathbb{R}$. En effet, la série de Fourier converge absolument, car $|a_n| \sim (1/\pi n^2)$ et la série de Riemann de terme général $1/n^2$ est convergente.
Effectivement, sur les figures de l'exemple 1 page 557, les approximations de Fourier semblent bien tendre en tout point vers la fonction.

Comportement en un point de discontinuité. Concernant la convergence de la série de Fourier en un point de discontinuité de la fonction, on a le résultat suivant.

Théorème de Dirichlet. *Si en un point x_0, la fonction f a une limite à gauche $\ell_- = \lim\limits_{\substack{x \to x_0 \\ x < x_0}} f(x)$ et une limite à droite $\ell_+ = \lim\limits_{\substack{x \to x_0 \\ x > x_0}} f(x)$, ainsi que des demi-dérivées à gauche et à droite, alors la somme de sa série de Fourier en x_0 est $\frac{1}{2}(\ell_+ + \ell_-)$.*

Exemple. La fonction g de l'exemple 2 page 557 possède au point π (par exemple), une limite à gauche $\ell_- = \pi$ et une limite à droite $\ell_+ = -\pi$: pour $x = \pi$, tous les termes de la série de Fourier sont nuls, donc aussi sa somme. La série de Fourier de g est $-2\sum_{n \geqslant 1} \dfrac{(-1)^n \sin nx}{n} = -2\sum_{n \geqslant 1} \dfrac{\sin n(x+\pi)}{n}$: elle converge pour tout x, car la série d'Abel $\sum_{n \geqslant 1} \dfrac{\sin ny}{n}$ est convergente quel que soit y (proposition page 530) ; mais si $y \neq 0$ modulo π, la série n'est pas absolument convergente.

4. Ondelettes de Haar

Nous avons vu que la décomposition de Fourier permet d'analyser des signaux périodiques de variable réelle. Voici une technique élémentaire pour traiter, de manière analogue, des signaux discrets, c'est-à-dire constitués d'une suite finie de valeurs réelles $f_0, f_1, \ldots, f_p$. Nous considérerons un tel signal discret comme une fonction constante par morceaux.

Donnons-nous un entier $N \geqslant 1$ et partageons l'intervalle $[0, 1[$ en parties égales à l'aide des 2^N points

$$x_0 = 0 \ , \ x_1 = \frac{1}{2^N} \ , \ldots \ , \ x_k = \frac{k}{2^N} \ , \ \ldots \ , \ x_{2^N-1} = \frac{2^N - 1}{2^N} = 1 - \frac{1}{2^N} .$$

Notons E_N l'espace vectoriel des fonctions à valeurs réelles, définies sur $[0, 1[$ et constantes sur chacun des intervalles $[x_k, x_{k+1}[$, où $0 \leqslant k \leqslant 2^N - 1$: pour $f \in E_N$, on a

$$f(x) = f_k \text{ si } x_k \leqslant x < x_{k+1}, \text{ pour tout entier } k \text{ tel que } 0 \leqslant k \leqslant 2^N - 1.$$

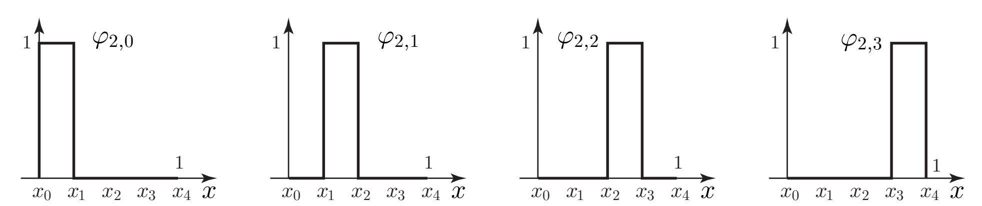
Pour toutes fonctions f et g appartenant à E_N, posons $f \cdot g = \int_0^1 f(t)g(t)\,dt$.
Cela définit un produit scalaire et E_N devient ainsi un espace euclidien (chapitre 7).

Une base de E_N. Considérons les fonctions $\varphi_{N,k}$ définies par

$$\varphi_{N,k}(x) = \begin{cases} 1 & \text{si } x_k \leqslant x < x_{k+1} \text{ et } 0 \leqslant k \leqslant 2^N - 1 \\ 0 & \text{sinon} \end{cases}$$

Si une fonction $f \in E_N$ vaut f_k sur $[x_k, x_{k+1}[$, alors pour tout $x \in [0, 1[$, on a $f(x) = \sum_{k=0}^{2^N-1} f_k \, \varphi_{N,k}(x)$, autrement dit toute fonction dans E_N s'écrit de manière unique comme la combinaison linéaire $f = \sum_{k=0}^{2^N-1} f_k \, \varphi_{N,k}$.

Les fonctions $\varphi_{N,k}$ forment une base de E_N. On a donc $\dim E_N = 2^N$.



Si k et ℓ sont des entiers différents, les intervalles $[x_k, x_{k+1}[$ et $[x_\ell, x_{\ell+1}[$ sont disjoints, donc pour x dans $[0,1[$, l'un au moins des nombres $\varphi_{N,k}(x)$ ou $\varphi_{N,\ell}(x)$ est nul. Ainsi la fonction produit $\varphi_{N,k}\varphi_{N,\ell}$ est nulle et par suite $\varphi_{N,k} \cdot \varphi_{N,\ell} = \int_0^1 \varphi_{N,k}(t)\varphi_{N,\ell}(t)\,dt = 0$.
D'autre part, $\int_0^1 \left[\varphi_{N,k}(t)\right]^2 dt = x_{k+1} - x_k = \dfrac{1}{2^N}$, d'où

$$\varphi_{N,k} \cdot \varphi_{N,\ell} = \begin{cases} 0 & \text{si } k \neq \ell \\ 1/2^N & \text{si } k = \ell \end{cases}$$

Pour $0 \leqslant k \leqslant 2^N - 1$, les fonctions $\varphi_{N,k}$ forment une base de E_N et sont deux à deux orthogonales.

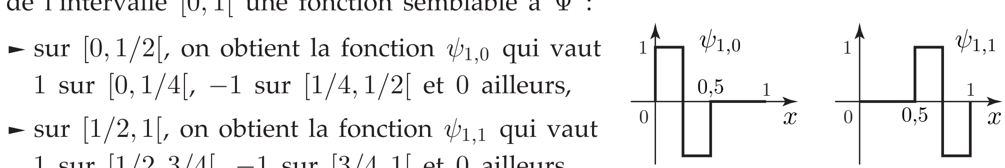
4.1 Génération des ondelettes

On part de l'ondelette-mère : c'est la fonction Ψ définie sur l'intervalle $[0,1[$ en posant

$$\Psi(x) = \begin{cases} 1 & \text{si } 0 \leqslant x < 1/2 \\ -1 & \text{si } 1/2 \leqslant x < 1 \end{cases}$$

Pour obtenir la première génération d'ondelettes, on reproduit dans chaque moitié de l'intervalle $[0,1[$ une fonction semblable à Ψ :

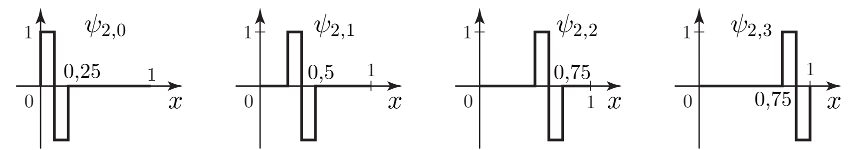
- sur $[0,1/2[$, on obtient la fonction $\psi_{1,0}$ qui vaut 1 sur $[0,1/4[$, -1 sur $[1/4,1/2[$ et 0 ailleurs,
- sur $[1/2,1[$, on obtient la fonction $\psi_{1,1}$ qui vaut 1 sur $[1/2,3/4[$, -1 sur $[3/4,1[$ et 0 ailleurs.



Chacune des fonctions $\psi_{1,0}$ et $\psi_{1,1}$ est constante sur les intervalles $[0,1/4[$, $[1/4,1/2[$, $[1/2,3/4[$ et $[3/4,1[$, donc appartient à l'espace E_2.
Poursuivons l'opération en faisant jouer à $\psi_{1,0}$ et $\psi_{1,1}$ le rôle de mère. La fonction $\psi_{1,0}$ donne naissance à deux fonctions $\psi_{2,0}$ et $\psi_{2,1}$:

- $\psi_{2,0}$ vaut 1 sur $[0,1/8[$ (première moitié de $[0,1/4[$), -1 sur $[1/8,1/4[$ et 0 ailleurs ;
- $\psi_{2,1}$ vaut 1 sur $[1/4,3/8[$ (première moitié de $[1/4,1/2[$), -1 sur $[3/8,1/2[$ et 0 ailleurs.

De même, la fonction $\psi_{1,1}$ donne naissance à $\psi_{2,1}$ sur l'intervalle $[1/2,3/4[$ et à $\psi_{2,2}$ sur $[3/4,1[$. Les quatre fonctions de cette deuxième génération appartiennent à l'espace E_3.



À la N-ième génération, on trouve 2^N fonctions $\psi_{N,k}$, où $0 \leqslant k \leqslant 2^N-1$, appelées *ondelettes de Haar* : la fonction $\psi_{N,k}$ vaut

- 1 sur la première moitié de l'intervalle $[k/2^N,(k+1)/2^N[$,
- -1 sur la seconde moitié,
- 0 en tout autre point de $[0,1[$.

La formule générale est

$$\psi_{N,k}(x) = \begin{cases} \Psi(2^N x-k) & \text{si } k/2^N \leqslant x < (k+1)/2^N \\ 0 & \text{sinon} \end{cases}$$

Puisque $\psi_{N,k}$ est constante sur tous les intervalles $[j/2^{N+1},(j+1)/2^{N+1}[$ pour $j=0,\dots,2^{N+1}-1$, c'est un élément de l'espace E_{N+1}. Remarquons qu'une ondelette est de moyenne nulle, car l'intégrale de $\psi_{N,k}$ entre 0 et 1 est égale à 0.

Terminologie. Si une fonction définie sur $[0,1[$ est nulle hors d'un intervalle $[a,b[$ inclus dans $[0,1[$, nous dirons qu'elle est *à support dans* $[a,b]$. Ainsi la fonction $\psi_{N,k}$ est à support dans $[k/2^N,(k+1)/2^N]$.
Supposons que u est une fonction à support dans $[a,b]$, que v est à support dans $[c,d]$ et que les intervalles ouverts $]a,b[$ et $]c,d[$ sont disjoints : on dit alors que u et v sont *à supports disjoints*. Dans ce cas, le produit $u(x)v(x)$ est nul sauf peut-être si $b=c=x$ et par suite $\int_0^1 u(t)v(t)\,dt = 0$.

Dans E_N, deux fonctions à supports disjoints sont orthogonales.

Puisque $\psi_{N,k}$ est à support dans $[x_k,x_{k+1}]$, $\psi_{N,k}$ et $\psi_{N,\ell}$ ont des supports disjoints si $k \neq \ell$. On en déduit :

pour $0 \leqslant k \leqslant 2^N-1$, les ondelettes $\psi_{N,k}$ sont deux à deux orthogonales dans l'espace E_{N+1}.

4.2 Décomposition en ondelettes

Rappelons que pour tout entier $k=0,1,\dots,2^N-1$, nous avons noté $\varphi_{N,k}$ la fonction appartenant à E_N qui est constante de valeur 1 sur $[x_k,x_{k+1}[$ et qui vaut 0 ailleurs. Ces 2^N fonctions sont deux à deux orthogonales.
Remarquons que tout élément de E_N est aussi élément de E_{N+1}, car une fonction constante sur un intervalle l'est aussi sur chaque moitié de cet intervalle. Ainsi E_N est un sous-espace vectoriel de E_{N+1}.

Une base de E_{N+1}. Considérons dans E_{N+1} la famille formée des $\varphi_{N,k}$ et des $\psi_{N,k}$, où $0 \leqslant k \leqslant 2^N - 1$ (pour $N = 2$, voir les figures pages 563 et 564). Calculons le produit scalaire de deux quelconques de ces vecteurs.

- Si $k \neq \ell$, alors $\varphi_{N,k} \cdot \varphi_{N,\ell} = \psi_{N,k} \cdot \psi_{N,\ell} = 0$, car les fonctions sont à supports disjoints.
- Pour la même raison, on a $\varphi_{N,k} \cdot \psi_{N,\ell} = 0$ si $k \neq \ell$.
- On a aussi $\varphi_{N,k} \cdot \psi_{N,k} = 0$.

En effet, puisque $\varphi_{N,k}(x)$ vaut 1 sur $[x_k, x_{k+1}[$ et 0 ailleurs, il vient

$$\int_0^1 \varphi_{N,k}(t)\psi_{N,k}(t)\,dt = \int_{x_k}^{x_{k+1}} \psi_{N,k}(t)\,dt = 0\,.$$

- $\varphi_{N,k} \cdot \varphi_{N,k} = \dfrac{1}{2^N} = \psi_{N,k} \cdot \psi_{N,k}$, car $\left|\varphi_{N,k}(x)\right|^2 = \left|\psi_{N,k}(x)\right|^2 = 1$ pour tout $x \in [x_k, x_{k+1}[$.

Ainsi, en mettant ensemble les $\varphi_{N,k}$ et les $\psi_{N,\ell}$, on obtient des vecteurs deux à deux orthogonaux dans E_{N+1}. Puisque aucune de ces fonctions n'est la fonction nulle, on en déduit que ce sont des vecteurs indépendants (proposition page 205). Le nombre de ces vecteurs est $2^N + 2^N = 2^{N+1}$ qui est la dimension de E_{N+1}, donc

la famille des $\varphi_{N,k}$ et $\psi_{N,k}$ est une base de E_{N+1} et ces vecteurs sont deux à deux orthogonaux.

Décomposition d'un signal. Rappelons que l'espace vectoriel E_{N+1} a aussi pour base les fonctions $\varphi_{N+1,j}$, où $0 \leqslant j \leqslant 2^{N+1} - 1$. L'entier N étant donné, posons pour simplifier $p = 2^N$,

$$\Phi_j = \varphi_{N+1,j} \text{ pour } 0 \leqslant j \leqslant 2p-1,\ \varphi_k = \varphi_{N,k} \text{ et } \psi_k = \psi_{N,k} \text{ pour } 0 \leqslant k \leqslant p-1.$$

Tout signal f appartenant à E_{N+1} se décompose de manière unique en

$$(*) \qquad f = \sum_{j=0}^{2p-1} f_j \Phi_j = \sum_{k=0}^{p-1} a_k \varphi_k + \sum_{k=0}^{p-1} b_k \psi_k$$

Entre une fonction Φ_j et l'une des fonctions φ_k ou ψ_k, les seuls produits scalaires non nuls sont

$$\Phi_{2k} \cdot \varphi_k = \Phi_{2k} \cdot \psi_k = \frac{1}{2p}\,, \quad \Phi_{2k+1} \cdot \varphi_k = \frac{1}{2p} \quad \text{et} \quad \Phi_{2k+1} \cdot \psi_k = \frac{-1}{2p}$$

En multipliant scalairement par Φ_{2k} dans $(*)$, il vient donc

$$f \cdot \Phi_{2k} = f_{2k}\,\Phi_{2k} \cdot \Phi_{2k} = a_k\,\Phi_{2k} \cdot \varphi_k + b_k\,\Phi_{2k} \cdot \psi_k = \frac{1}{2p}a_k + \frac{1}{2p}b_k$$

De même, en multipliant scalairement par Φ_{2k+1}, on obtient

$$f \cdot \Phi_{2k+1} = f_{2k+1}\,\Phi_{2k+1} \cdot \Phi_{2k+1} = a_k\,\Phi_{2k+1} \cdot \varphi_k + b_k\,\Phi_{2k+1} \cdot \psi_k = \frac{1}{2p}a_k - \frac{1}{2p}b_k$$

Puisque $\Phi_j \cdot \Phi_j = \frac{1}{2p}$, on a simplement les relations $f_{2k} = a_k + b_k$ et $f_{2k+1} = a_k - b_k$, d'où

$$a_k = \frac{1}{2} f_{2k} + \frac{1}{2} f_{2k+1} \quad \text{et} \quad b_k = \frac{1}{2} f_{2k} - \frac{1}{2} f_{2k+1}, \quad \text{pour } 0 \leqslant k \leqslant p-1.$$

Ce sont les formules du changement de base. La fonction $P_N(f) = \sum_{k=0}^{p-1} a_k \varphi_k$ appartient à E_N et $f - P_N(f) = \sum_{k=0}^{p-1} b_k \psi_k$ est orthogonale à E_N. D'après le théorème de la projection (page 209), on en déduit :

$$\sum_{k=0}^{p-1} a_k \varphi_k \textit{ est le projeté orthogonal de } f \textit{ sur le sous-espace } E_N.$$

La fonction f est constante sur des intervalles de longueur $1/2^{N+1}$: si l'on veut approcher f par des fonctions constantes sur des intervalles de longueur double, la meilleure approximation au sens de la norme dans E_{N+1} est $P_N(f) = \sum_{k=0}^{p-1} a_k \varphi_k$.

On peut bien sûr poursuivre la décomposition en projetant $P_N(f)$ sur E_{N-1} et ainsi de suite. Cela conduit à calculer les coordonnées de f dans la base de E_{N+1} formée de la constante 1, de $\psi_0 = \Psi$ et de tous les $\psi_{n,k}$, où $n = 1, 2, \ldots, N$ et $0 \leqslant k \leqslant 2^n - 1$ (le nombre de ces fonctions est bien $1+1+2+2^2+\cdots+2^N = 2^{N+1} = \dim E_{N+1}$).

On peut définir d'autres types d'ondelettes en choisissant comme ondelette-mère une courbe convenable. Cela permet, comme la décomposition de Fourier, d'analyser aussi des signaux continus.

Exemple. Prenons $N = 2$. Pour tout signal discret $f = [f_j]_{0 \leqslant j \leqslant 7}$ appartenant à E_3, les coordonnées a_0, a_1, a_2, a_3 et b_0, b_1, b_2, b_3 sont données par le produit matriciel

$$[a_0\, a_1\, a_2\, a_3\, b_0\, b_1\, b_2\, b_3] = [f_0\, f_1\, f_2\, f_3\, f_4\, f_5\, f_6\, f_7] \begin{bmatrix} 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & -1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & -1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & -1/2 \end{bmatrix}$$

Écrivons cela sous la forme condensée $[a\ b] = [f]H$, où l'on a appelé H la matrice 8×8, posé $a = [a_0\, a_1\, a_2\, a_3]$, $b = [b_0\, b_1\, b_2\, b_3]$ et noté $[f]$ le vecteur-ligne constitué des valeurs du signal.

Plus généralement, introduisons la matrice carrée $H = [h_{ij}]$ de taille $p = 2^{N+1}$ définie par

$$h_{ij} = \begin{cases} 1/2 & \text{si } j \leqslant p/2 \text{ et } i = 2j-1 \text{ ou } 2j \\ 1/2 & \text{si } j > p/2 \text{ et } i = 2j-p-1 \\ -1/2 & \text{si } j > p/2 \text{ et } i = 2j-p \\ 0 & \text{sinon} \end{cases}$$

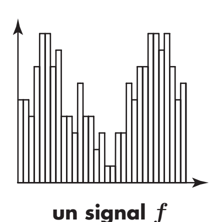
En employant les notations introduites dans l'exemple, les formules qui permettent de calculer les a_k et les b_k s'expriment par la relation matricielle

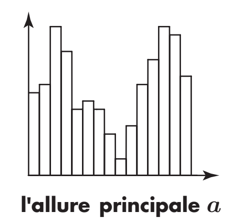
$$[\,a\ \ b\,] = [f]\,H\ ,\quad \text{où le signal } f \text{ est de taille } p.$$

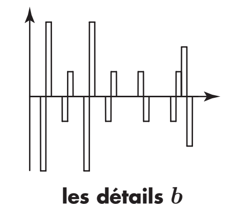
Puisque H est une matrice de changement de base, elle est inversible.

Interprétation de la décomposition

Soit $f = [f_0, f_1, \ldots, f_{2^N-1}]$ un signal discret de longueur 2^N. Si l'on veut approcher f (au sens de la distance dans E_N) par un signal de longueur moitié, le meilleur choix est le projeté $a = [a_0, a_1, \ldots, a_{2^{N-1}-1}]$, où $a_k = \frac{1}{2}f_{2k} + \frac{1}{2}f_{2k+1}$.
Puisque a_k est la moyenne de deux valeurs consécutives de f, le signal a est plus régulier que f : des détails de f n'y apparaissent pas, mais on y retrouve l'allure principale du signal initial.
L'information de détail est exprimée dans le signal b. Si f est assez régulier, des valeurs consécutives f_{2k} et f_{2k+1} sont voisines et les termes b_k sont petits.
La figure de gauche ci-dessous montre le graphe d'un signal f échantillonné sur $2^5 = 32$ points : la fonction est constante sur chacun des sous-intervalles de partage. Au centre, on voit le graphe de la partie principale : c'est une fonction constante sur seize sous-intervalles de longueur double des précédents. Les détails sont représentés par le graphe de droite, où la fonction est une combinaison des seize ondelettes $\psi_{4,k}$, $k = 0, \ldots, 15$. Il y a plusieurs intervalles où cette fonction de détail est nulle, correspondant aux valeurs de k telles que $b_k = 0$.


un signal f


l'allure principale a


les détails b

4.3 Application à la compression d'images

On numérise une image en la découpant en $2^N \times 2^N$ pixels répartis en 2^N lignes, chaque ligne comprenant 2^N pixels. Dans le cas d'une image en noir et blanc, on associe à chaque pixel un niveau de gris qu'on peut repérer sur une échelle convenable (formée par exemple d'entiers entre 0 et 255 ou par des nombres entre 0 et 1). L'image est ainsi définie par une matrice carrée de taille 2^N. Pour une image en couleurs, il suffit de considérer trois images formées des niveaux dans les couleurs primaires (rouge, vert et bleu). Supposons désormais que notre image est en noir et blanc.

Traitement par ondelettes. Soit M la matrice qui définit l'image. Chaque ligne et chaque colonne est un signal de longueur 2^N que l'on peut décomposer comme

précédemment. Une ligne L se décompose en LH et une colonne C en $({}^tH)C$, donc

la matrice M se décompose en la matrice $M' = ({}^tH)MH$.

À partir de M', on retrouve M par le produit matriciel $({}^tH)^{-1}M'(H^{-1}) = M$.
Posons $M' = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, où A, B, C, D sont des matrices carrées de taille 2^{N-1}.

- L'allure principale de l'image est donnée par la matrice A et l'information de détail se trouve dans les matrices B, C et D.
- L'image se présente en général sous forme de zones relativement homogènes où l'intensité varie peu, séparées par des lignes de contour où la variation d'intensité est brutale. Dans une zone donnée, chaque pixel a donc des voisins à peu près de même valeur. Il en résulte que dans les matrices B, C et D, la plupart des coefficients sont petits, voire nuls, car dans M, la différence entre deux coefficients successifs sur une même ligne ou sur une même colonne est petite : on dit que ce sont les *coefficients de détail*.

Principe de la compression

A priori, l'image est définie par $2^N \times 2^N$ nombres. Mais donnons-nous une tolérance $\varepsilon > 0$ et remplaçons par 0 tous les coefficients de détail dont la valeur absolue est moindre que ε. La matrice M' devient

$$M'_\varepsilon = \begin{bmatrix} A & B_\varepsilon \\ C_\varepsilon & D_\varepsilon \end{bmatrix}$$

où les matrices B_ε, C_ε et D_ε contiennent beaucoup de zéros : pour cette raison, on dit que ce sont des *matrices creuses*.

Pour définir une matrice creuse, il suffit de se donner les indices et les valeurs des coefficients non nuls, en convenant que les autres coefficients sont nuls : c'est une donnée beaucoup moins volumineuse que la matrice elle-même, de sorte qu'une fois numérisée, cette information génère un fichier de données plus petit que si l'on y avait indiqué la valeur de chacun des coefficients de M.

La matrice creuse M'_ε contient les coefficients principaux et peut se coder de manière beaucoup plus économique que M.

À partir de la matrice M'_ε, on peut reconstituer une image un peu dégradée en effectuant le produit $M_\varepsilon = ({}^tH)^{-1}M'_\varepsilon(H^{-1})$ et en interprétant les coefficients de M_ε comme des valeurs de gris pour les pixels.
En acceptant une image de moins bonne définition, il est donc possible de diminuer de façon significative le volume de l'information et de gagner ainsi en rapidité de transmission. C'est sur ce principe que fonctionnent les logiciels de compression d'images.

Algorithme de compression

initialisations :

$M \leftarrow$ la matrice carrée de taille $p{=}2^N$ correspondant à l'image initiale numérisée
$\varepsilon > 0 \ \leftarrow$ une tolérance

i) $[m'_{ij}] \leftarrow ({}^tH)MH$

ii) *boucles :* pour $i = 1, \ldots, p$, pour $j = 1, \ldots, p$,

si $i \geqslant (p/2)+1$ ou si $j \geqslant (p/2)+1$, alors

si $|m'_{ij}| < \varepsilon$, faire $m'_{ij} \leftarrow 0$.

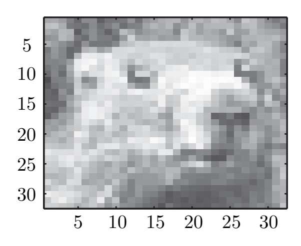
iii) *fin :* $M'' \ \leftarrow \ ({}^tH)^{-1}[m'_{ij}](H^{-1})$.

À la fin de l'algorithme,

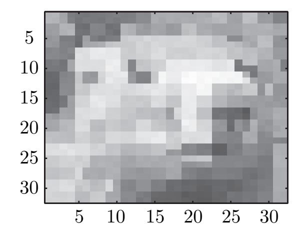
- la matrice $M' = [m'_{ij}]$ contient une version compressée de l'image : c'est la matrice qu'on peut transmettre par voie informatique ;
- la matrice M'' représente une image dégradée avec tolérance ε.

Exemple. On a codé l'image d'une tête de marmotte au moyen d'une matrice carrée M de taille 32. Les degrés de gris ont été repérés sur une échelle de 0 à 1 et l'on a choisi la tolérance $\varepsilon = 0{,}1$.

Dans la matrice $M'_\varepsilon = \begin{bmatrix} A & B_\varepsilon \\ C_\varepsilon & D_\varepsilon \end{bmatrix}$, les sous-matrices B_ε, C_ε et D_ε sont bien creuses : dans B_ε, il n'y a environ que 7,4% de coefficients non nuls, dans C_ε, il y en a 4% et $D_\varepsilon = 0$. Au total, la matrice M'_ε n'a que 27,8% de coefficients non nuls (dont les 25% de coefficients principaux qui n'ont pas changé). On voit que, malgré la compression, l'image se dégrade peu.



avant compression



après compression

Exercices

1. Calcul approché de la somme d'une série Dans l'exemple page 530, nous avons montré que si $a > 0$, alors $\sum_{k=p}^{+\infty} \frac{1}{a^2+k^2} \leqslant \frac{1}{a^2+p^2} + \frac{1}{a} \operatorname{Arc\,tan} \frac{a}{p}$ pour tout entier $p \geqslant 1$. Montrer que pour tout $x \geqslant 0$, on a l'encadrement $0 \leqslant \operatorname{Arc\,tan} x \leqslant x$. En déduire $\sum_{k=p}^{+\infty} \frac{1}{k^2+p^2} \leqslant 2/p$ puis, au moyen d'un ordinateur, une valeur approchée à 2.10^{-2} près de la somme $\sum_{k=1}^{+\infty} \frac{1}{k^2+p^2}$.

@ **2. Développement en série entière et suite récurrente linéaire.** Soient p et q des nombres. On suppose $q \neq 0$ et pour tout x, on pose $f(x) = \frac{1}{q+px+x^2}$.

a) Montrer que si $|x|$ est assez petit, $f(x)$ est bien défini.

b) Montrer que pour une série entière $\sum a_k x^k$, on a $(q+px+x^2)\sum a_k x^k = 1$ si et seulement si $qa_0 = 1$, $qa_1 + pa_0 = 0$ et $qa_{k+2} + pa_{k+1} + a_k = 0$ quel que soit $k \geqslant 0$. En déduire que pour $|x|$ assez petit, on a le développement en série entière $\frac{1}{q+px+x^2} = \sum_{k=0}^{+\infty} a_k x^k$.

c) Montrer que le début du développement en série entière de $f(x)$ est
$$\frac{1}{q+px+x^2} = \frac{1}{q} - \frac{p}{q^2}x + \frac{-q+p^2}{q^3}x^2 - \frac{p(-2q+p^2)}{q^4}x^3 + \frac{q^2-3qp^2+p^4}{q^5}x^4 + \cdots$$

3. Les nombres de Fibonacci. On définit les nombres de Fibonacci F_n en posant $F_0 = 0$, $F_1 = 1$ et $F_{k+2} = F_{k+1} + F_k$ pour tout entier $k \geqslant 0$. Si $k > 0$, F_k est un entier positif.

a) Montrer que pour $|x|$ assez petit, le développement en série entière de $\frac{x}{1-x-x^2}$ est $\sum_{k=0}^{+\infty} F_k x^k$ (multiplier cette série entière par $1-x-x^2$).

b) Calculer les racines de $1-x-x^2$ et montrer que l'on a $x^2+x-1 = (x+\alpha)(x+\beta)$, où $\alpha = \frac{1+\sqrt{5}}{2}$ et $\beta = \frac{1-\sqrt{5}}{2}$. Vérifier les relations $\alpha+\beta = 1$ et $\alpha\beta = -1$.

c) Vérifier l'égalité $\frac{1}{1-x-x^2} = \frac{1}{\alpha-\beta}\left[\frac{1}{\alpha+x} - \frac{1}{\beta+x}\right]$. En déduire le développement en série entière $\frac{1}{1-x-x^2} = \frac{1}{\alpha-\beta}\sum_{k=0}^{+\infty}(\alpha^{k+1}-\beta^{k+1})x^k$ (utiliser le développement en série entière de $\frac{1}{a-z}$). Montrer finalement que l'on a $\frac{x}{1-x-x^2} = \frac{1}{\alpha-\beta}\sum_{k=1}^{+\infty}(\alpha^k-\beta^k)x^k$ pour $|x| < |\beta|$.

d) En déduire que pour tout entier $k \geqslant 0$, on a l'égalité $F_k = \frac{(1+\sqrt{5})^k - (1-\sqrt{5})^k}{2^k\sqrt{5}}$.

La fraction de droite est donc un nombre entier, car les nombres de Fibonacci sont par définition des entiers.

4. Utilisation d'une série génératrice. Soient p, q et r des entiers strictement positifs. Étant donné un entier $n \geqslant 0$, on note u_n le nombre de solutions (x, y, z) de l'équation $px + qy + rz = n$, où les inconnues x, y, z sont des entiers positifs ou nuls.

a) Écrire le développement en série entière de la fonction $t \mapsto \frac{1}{1-t^p}$. Quel est le rayon de convergence ?

b) Montrer que u_n est le coefficient de t^n dans le développement en série entière de $\frac{1}{(1-t^p)(1-t^q)(1-t^r)}$.

c) On prend $p=2$, $q=3$, $r=5$. Vérifier que l'on a par exemple $u_{20}=11$, $u_{20+k}=10+k$ pour $1 \leqslant k \leqslant 9$ et $u_{30} = u_{31} = 21$.

@ **5. Un problème de file d'attente.** On considère un guichet où des clients se présentent à des instants numérotés $0, 1, 2, \ldots$ Les arrivées sont aléatoires, mais on suppose qu'entre deux instants consécutifs $n-1$ et n, il y a toujours la même probabilité p qu'un client se présente.

On appelle durée de service d'un client le temps passé au guichet à le servir. On

suppose que ces durées sont des variables aléatoires D indépendantes qui suivent une même loi de Poisson, c'est-à-dire de probabilité

$$P(D = n) = e^{-\lambda}\,\frac{\lambda^n}{n!}\ \text{, pour tout entier } n \geqslant 0,$$

où λ est un paramètre strictement positif.
La première vague de clients est formée des personnes arrivées pendant le temps de service du premier client et la $(k+1)$-ième vague est formée des personnes arrivées pendant la durée des services des clients de la k-ième vague.
Notons N_k le nombre de clients de la k-ième vague, en convenant $N_0 = 1$.

a) Soit D la durée de service du premier client. Montrer que la probabilité pour que $N_1 = k$ sachant que $D = n$ est $P(N_1 = k \mid D = n) = \binom{n}{k} p^k (1-p)^{n-k}$ (rappelons que le coefficient binomial $\binom{n}{k}$ est nul si $k > n$). En déduire $P(N_1 = k) = \sum_{n=0}^{+\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda}\,\frac{\lambda^n}{n!}$.

b) En écrivant la somme ci-dessus sous la forme $e^{-\lambda}\,\frac{\lambda^k p^k}{k!} \sum_{n=k}^{+\infty} \frac{(1-p)^{n-k}\lambda^{n-k}}{(n-k)!}$, montrer que $P(N_1 = k) = e^{-\lambda p}\,\frac{(\lambda p)^k}{k!}$. En déduire que l'espérance de N_1 est λp.

Posons $p_k = P(N_k = 0)$. S'il n'y a personne à la k-ième vague ($N_k = 0$), il n'y a personne non plus à la vague suivante ($N_{k+1} = 0$) : on a donc $p_k \leqslant p_{k+1}$. La suite des probabilités p_k est croissante et majorée par 1, donc les p_k ont une limite. La réunion de tous les événements $N_k = 0$ pour $k = 1, 2, \ldots$ est l'événement « la file d'attente s'achève en un temps fini » : la probabilité de cet événement est donc $\lim\limits_{k\to+\infty} p_k$.

c) Supposons que la première vague contienne j clients. Dire que la $(k+1)$-ième vague est vide signifie que lorsqu'on fait jouer à chacune de ces j personnes le rôle d'un client initial, la k-ième vague est vide : la probabilité pour que $N_{k+1} = 0$ sachant que $N_1 = j$ est donc $\left[P(N_k = 0)\right]^j = (p_k)^j$. En utilisant l'égalité $P(N_{k+1} = 0) = \sum_{j=0}^{+\infty} P(N_{k+1} = 0 \mid N_1 = j)\,P(N_1 = j)$, montrer que pour tout $k \geqslant 0$, on a $p_{k+1} = \sum_{j=0}^{+\infty} (p_k)^j e^{-\lambda p}\,\frac{(\lambda p)^j}{j!} = e^{\lambda p(p_k - 1)}$, où $p_0 = 0$.

d) Reportons-nous à l'exercice **3** page 337 où l'on a étudié, en fonction du paramètre $a > 0$, la suite (u_k) définie par l'itération $u_{k+1} = e^{a(u_k - 1)}$ et la valeur initiale $u_0 = 0$. En utilisant les résultats de cet exercice,

(i) justifier l'affirmation suivante : si $\lambda p \leqslant 1$, il est presque certain qu'au bout d'un temps fini, le guichet se libère ;

(ii) vérifier le tableau ci-dessous où figurent, pour quelques valeurs de p et de λ, des estimations de la probabilité pour que le guichet se libère au bout d'un temps fini.

	$p = 0{,}3$	$p = 0{,}5$	$p = 0{,}75$
$\lambda = 4$	0,68	0,20	0,06
$\lambda = 5$	0,42	0,10	0,02

@ **6. Une propriété des polynômes de Legendre.** Pour tout nombre z, on définit la fonction $V(h) = (1-2zh+h^2)^{-1/2}$, où la variable h est réelle.

a) Montrer que $(1-2zh+h^2)\dfrac{dV}{dh} = (z-h)V$.

b) La fonction V est développable en série entière : montrer que les coefficients sont des polynômes en z, autrement dit que si $|h|$ est assez petit, alors $V(h) = Q_0(z) + Q_1(z)h + Q_2(z)h^2 + \cdots + Q_k(z)h^k + \cdots$, où $Q_0, Q_1, Q_2, \ldots$ sont des polynômes. Calculer Q_0 et Q_1 et vérifier que l'on a $Q_2(z) = (3/2)z^2-(1/2)$.

c) Montrer que le coefficient de h^n dans $(1-2zh+h^2)\dfrac{dV}{dh}$ est

$$(n+1)Q_{n+1} - 2nzQ_n + (n-1)Q_{n-1}\,.$$

Déduire de **(a)** que les polynômes Q_n satisfont la même relation de récurrence que les polynômes de Legendre : $(n+1)Q_{n+1} = (2n+1)zQ_n - nQ_{n-1}$.

d) Soit z un nombre donné. Montrer que si $|h|$ assez petit, alors

$$(1-2zh+h^2)^{-1/2} = P_0(z) + P_1(z)h + P_2(z)h^2 + \cdots + P_k(z)h^k + \cdots$$

où $P_0, P_1, P_2, \ldots$ sont les polynômes de Legendre.

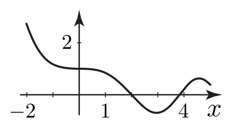
@ **7. L'équation différentielle d'Airy.** Il s'agit de l'équation $y'' + xy = 0$ (exercice **11** page 158 et exercice **5** page 473). Cherchons des solutions sous la forme $y(x) = \sum_{n=0}^{+\infty} a_n x^n$.

a) Montrer que les a_n satisfont la relation $(n+2)(n+3)a_{n+3} + a_n = 0$ pour tout $n \geqslant 0$.

b) En déduire que la solution telle que $y(0) = 1$ et $y'(0) = 0$ est la fonction $A(x) = 1+\sum_{n=1}^{+\infty}(-1)^n \dfrac{x^{3n}}{2\cdot 3\,\cdot\, 5\cdot 6\cdots(3n-1)(3n)}$ et que cette solution est définie sur $\mathbb{R}$.

Les premiers termes de la série sont $1-\dfrac{x^3}{6}+\dfrac{x^6}{180}-\cdots$

c) Posons $A_N(x) = 1+\sum_{n=1}^{N}(-1)^n \dfrac{x^{3n}}{2\cdot 3\,\cdot\, 5\cdot 6\cdots(3n-1)(3n)}$, où $N \geqslant 1$. Faire dessiner par ordinateur le graphe de la fonction A_9 sur l'intervalle $[-2, 5]$. Montrer que si $a > 0$, alors pour tout $x \in [0, a]$, on a $\big|A(x)-A_N(x)\big| \leqslant \dfrac{a^{3N+3}}{2\cdot 3\,\cdot\, 5\cdot 6\cdots(3N+2)(3N+3)}$. Après avoir trouvé une majoration convenable de l'écart $\big|A(-2)-A_9(-2)\big|$, vérifier que pour $-2 \leqslant x \leqslant 5$, on a $\big|A(x)-A_9(x)\big| \leqslant 6.10^{-2}$.



graphe de $A(x)$ entre -2 et 5

8. Soit f un signal périodique de période T. On décale f dans le temps d'une quantité a en posant $f_a(x) = f(x-a)$. Montrer que pour tout entier n, on a $c_n(f_a) = e^{-n\boldsymbol{i}\omega a}c_n(f)$.

9. Soit g la fonction périodique de période 2π telle que $g(x) = x$ si $-\pi < x \leqslant \pi$ (exemple **2** page 557). Utiliser la série de Fourier de g pour montrer que lorsque $0 < x < \pi$, on a l'égalité $\sum_{n=1}^{+\infty} \dfrac{\sin nx}{n} = \dfrac{\pi}{2} - \dfrac{x}{2}$.

@ **10. Calcul d'une série de Fourier.** Posons $u(x) = \dfrac{\pi}{8}x(\pi-x)$ pour $0 \leqslant x < \pi$. On prolonge u en une fonction f impaire et périodique de période 2π.

a) Montrer que pour tout entier n, on a $\int_{-\pi}^{\pi} f(t)\sin nt\,dt = 2\int_0^{\pi} u(t)\sin nt\,dt$.

b) Montrer que $\int_0^{\pi} u(t)\sin nt\,dt$ vaut 0 si n est pair et $\frac{\pi}{2n^3}$ si n est impair. En déduire que pour $0 \leqslant x < \pi$, on a l'égalité $\frac{\pi}{8}x(\pi - x) = \sum_{n=0}^{+\infty} \frac{\sin(2n+1)x}{(2n+1)^3}$.

c) Montrer les égalités $\sum_{n=0}^{+\infty} \frac{(-1)^n}{(2n+1)^3} = \frac{\pi^3}{32}$ et $\sum_{n=0}^{+\infty} \frac{1}{(2n+1)^6} = \frac{\pi^6}{960}$.

11. Oscillations dans un bassin rectangulaire. On laisse osciller librement la surface de l'eau dans un bassin rectangulaire de longueur a, de largeur b et de profondeur h. Choisissons des axes Ox et Oy comme sur la figure ci-contre. L'élévation de l'eau au dessus du niveau de repos est une fonction $Z(x, y, t) = \cos(\omega t + \varphi)u(x, y)$ solution de l'équation

$$\frac{\partial^2 Z}{\partial t^2} = c^2\,\Delta Z\ ,\ \text{où } c = \sqrt{gh} \text{ et } \Delta Z = \frac{\partial^2 Z}{\partial x^2} + \frac{\partial^2 Z}{\partial y^2}.$$

De plus, en un point de la paroi, il n'y a pas de déplacement dans la direction perpendiculaire à celle-ci : on a donc les conditions aux bords $\frac{\partial u}{\partial x} = 0$ si $x = 0$ ou $x = a$, et $\frac{\partial u}{\partial y} = 0$ si $y = 0$ ou $y = b$.

a) Montrer que l'équation s'écrit $\Delta u + \frac{\omega^2}{c^2}u = 0$.

b) Cherchons une solution sous la forme $u(x, y) = A\cos(\alpha x + \varphi_1)\cos(\beta y + \varphi_2)$, où A est une constante. Montrer que les conditions aux bords imposent $u(x, y) = A_{k,\ell}\cos\left(\frac{k\pi x}{a}\right)\cos\left(\frac{\ell\pi y}{b}\right)$, avec k et ℓ des entiers.

c) Calculer Δu et en utilisant **(a)**, montrer que $\omega^2 = \pi^2 c^2\left[\frac{k^2}{a^2} + \frac{\ell^2}{b^2}\right]$. En déduire la solution $Z_{k,\ell} = A_{k,\ell}\cos\left(\frac{k\pi x}{a}\right)\cos\left(\frac{\ell\pi y}{b}\right)\cos(\omega_{k,\ell}t + \varphi_{k,\ell})$, où k et ℓ sont des entiers et où $\omega_{k,\ell} = \pi c\left[\frac{k^2}{a^2} + \frac{\ell^2}{b^2}\right]^{1/2}$.

d) Expliquer le premier mode d'oscillation de la surface de l'eau ($a > b$, $k = 1$ et $\ell = 0$) : il y a une seule vague longitudinale qui va et vient, avec un noeud à mi-longueur de bassin ($x = a/2$).

On peut voir un autre mode d'oscillation sur les figures 2 et 3 page 360.

En décomposant la fonction $u(x, y)$ en double-série de Fourier, on montre que la solution générale est de la forme $\sum_{k,\ell} Z_{k,\ell}$, la sommation se faisant sur tous les couples d'entiers k, ℓ.

@12. Calcul d'une solution périodique d'équation différentielle. On considère l'équation différentielle linéaire $y'' - 2y' + 2y = |\sin(t/2)|$. Le second membre étant 2π-périodique, cherchons une solution 2π-périodique $t \mapsto f(t)$ sous la forme de son développement de Fourier $f(t) = \sum_{n\in\mathbb{Z}} c_n e^{\boldsymbol{i}nt}$.

a) Montrer que pour tout t, on a $|\sin(t/2)| = \dfrac{2}{\pi} - \dfrac{4}{\pi}\displaystyle\sum_{n=1}^{+\infty}\frac{\cos nt}{4n^2-1}$.

b) En déduire que les c_n doivent satisfaire les relations $c_0 = \dfrac{1}{\pi}$ et $(n^2-2+2n\boldsymbol{i})c_n = \dfrac{2}{\pi}\dfrac{1}{4n^2-1}$ pour $n \in \mathbb{Z}$, $n \neq 0$. Calculer la partie réelle et la partie imaginaire de c_n.
Admettons les résultats suivants :

- si $n^3|c_n|$ a une limite finie quand n tend vers $\pm\infty$, alors f a une dérivée continue ;
- si $n^4|c_n|$ a une limite finie quand n tend vers $\pm\infty$, alors f a une dérivée seconde continue.

d) Montrer que la fonction $f(t) = \dfrac{1}{\pi} + \dfrac{4}{\pi}\displaystyle\sum_{n=1}^{+\infty}\frac{(n^2-2)\cos nt}{(4n^2-1)(n^4+4)} + \frac{4}{\pi}\sum_{n=1}^{+\infty}\frac{2n\sin nt}{(4n^2-1)(n^4+4)}$ est solution de l'équation différentielle.

@13. La formule de Poisson. Soit $h : \mathbb{R} \to \mathbb{R}$ une fonction continue. Supposons que lorsque x tend vers $\pm\infty$, $h(x)$ est négligeable devant $1/|x|^\alpha$, où $\alpha > 1$.

a) Soient $a > 0$, $x \in [-a, a]$ et $|n| > a$. Montrer que $|x+n| \geqslant |n|-a$. En déduire que les nombres $n^\alpha\big|h(x+n)\big|$ sont majorés quand x parcourt $[-a, a]$ et quand n parcourt $\mathbb{Z}$. Conclure qu'il existe une série $\sum a_n = \sum M/n^\alpha$ convergente telle que $\big|h(x+n)\big| \leqslant a_n$ pour tout $n \in \mathbb{Z}$ et tout $x \in [-a, a]$. La série $\sum h(x+n)$ est donc absolument convergente quel que soit $x \in \mathbb{R}$.

b) Pour tout $x \in \mathbb{R}$, posons $f(x) = \sum_{n=-\infty}^{+\infty} h(x+n)$.

i) Montrer que la fonction f est périodique de période 1.

ii) Pour tout entier relatif p, écrivons $c_p(f) = u_N + A_N$, où

$$U_N = \int_0^1 \left[\sum_{n=-N}^{N} h(t+n)\right] e^{-2\boldsymbol{i}\pi pt}\,dt \text{ et } A_N = \int_0^1 \left[\sum_{|n|>N} h(t+n)\right] e^{-2\boldsymbol{i}\pi pt}\,dt.$$

Montrer que $|A_N| \leqslant \sum_{|n|>N} a_n$. En déduire $\lim\limits_{N\to+\infty} |A_N| = 0$.

iii) Montrer que $\displaystyle\int_0^1 h(t+n)e^{-2\boldsymbol{i}\pi pt}\,dt = \int_n^{n+1} h(s)e^{-2\boldsymbol{i}\pi ps}\,ds$. En déduire $U_N = \displaystyle\int_{-N}^{N+1} h(s)e^{-2\boldsymbol{i}\pi ps}\,ds$.

(iv) En utilisant l'hypothèse faite sur h, montrer que l'intégrale généralisée $\displaystyle\int_{-\infty}^{+\infty} h(s)e^{-2\boldsymbol{i}\pi ps}\,ds$ existe.

c) Pour tout entier $p \in \mathbb{Z}$, posons $\hat{h}(p) = \displaystyle\int_{-\infty}^{+\infty} h(s)e^{-2\boldsymbol{i}\pi ps}\,ds$. Déduire de **(b) (ii)** et **(iii)** que l'on a $c_p(f) = \hat{h}(p)$.

d) Supposons de plus que la série $\sum\big|\hat{h}(p)\big|$ est convergente. Montrer que pour tout $x \in \mathbb{R}$, on a l'égalité $\sum_{n=-\infty}^{+\infty} h(x+n) = \sum_{n=-\infty}^{+\infty} \hat{h}(n)e^{2\boldsymbol{i}\pi nx}$. En déduire *la formule de Poisson* : $\sum_{n=-\infty}^{+\infty} h(n) = \sum_{n=-\infty}^{+\infty} \hat{h}(n)$.

@14. Utilisons les notations du paragraphe sur les ondelettes (page 567).
Soit $f \in E_N$ et soit $[a\,b]$ la décomposition de f en partie principale a et détails b.

a) Montrer que $\|f\|^2 = \|a\|^2 + \|b\|^2$.

b) Montrer que les signaux f et a ont la même moyenne, autrement dit $\int_0^1 f(t)\,dt = \int_0^1 a(t)\,dt$.

Annexes

1. Fonction de Gauss

Table de valeurs de la fonction de Gauss $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$.

x	0,0	0,1	0,2	0,3	0,4	0, 5	0,6	0,7	0,8	0,9
$\Phi(x)$	0,5000	0,5398	0,5792	0,6179	0,6554	0,6914	0,7257	0,7580	0,7881	0,8159
x	1,0	1,1	1,2	1,3	1,4	1, 5	1,6	1,7	1,8	1,9
$\Phi(x)$	0,8413	0,8643	0,8849	0,9032	0,9192	0,9331	0,9452	0,9554	0,9640	0,9712
x	2,0	2,1	2,2	2,3	2,4	2, 5	2,6	2,7	2,8	2,9
$\Phi(x)$	0,9772	0,9821	0,9861	0,9892	0,9918	0,9937	0,9953	0,9965	0,9974	0,9981

Pour les valeurs négatives de la variable, utiliser la relation $\Phi(-x) = 1 - \Phi(x)$.

2. Fonctions de Bessel

Voici les valeurs des premiers zéros $x_k > 0$ de la fonction de Bessel J_0 et des premiers zéros $y_k > 0$ de la fonction de Bessel $J_1 = -J_0'$: on a $J_0(x_k) = 0 = J_1(y_k)$.

k	1	2	3	4	5	6	7	8	9	10
x_k	2,4048	5,5200	8,6537	11,7915	14,9309	18,0710	21,2116	24,3524	27,4934	30,6346
y_k	3,8317	7,0155	10,1734	13,3236	16,4706	19,6158	22,7600	25,9036	29,0468	32,1896

3. Analyse de données

Le tableau ci-dessous présente des données sociologiques recueillies dans dix-huit villes françaises et complétées par des résultats sur les élections municipales des 11 et 18 mars 2001 (source : Le Monde).

	poptot	popetr	logsoc	chomage	txhab	revenu	votants	maj
Poitiers	83507	3216	24,00	9,70	21,96	43082	25274	53,00
Reims	187181	10117	45,00	10,90	10,89	43924	49318	51,58
Marseille	797491	54355	17,00	17,70	21,93	43801	214637	48,57
Le Havre	190924	8208	33,40	14,40	17,67	41060	63462	57,87
Besançon	117691	7947	26,80	8,00	18,83	43751	33191	55,30
Strasbourg	263941	34138	23,60	6,10	14,26	48878	67987	50,85
Nice	343123	30914	8,00	10,90	18,96	48549	104373	44,48
Lyon	445274	35583	15,00	12,60	11,62	56860	134347	48,56
Amiens	135449	6180	34,16	14,00	17,97	39935	41854	52,06
St Denis Réunion	131649	1394	35,00	30,00	13,43	35775	40348	51,11
Nancy	103552	6133	18,00	10,10	11,56	49716	26566	50,81
Toulouse	390301	28073	17,30	12,50	17,99	45941	122901	55,13
Brest	149649	2997	18,49	9,90	15,42	41559	44439	57,55
Boulogne Bill.	106316	12600	8,32	6,10	9,27	94074	29682	66,08
Evry	49397	6472	47,00	9,70	17,62	41568	10142	44,17
Versailles	85761	4732	15,70	5,71	8,95	82208	28306	50,43
Neuilly	59874	6584	2,50	7,20	3,94	196642	17759	76,83
St Denis	85994	22535	49,04	20,40	9,17	36534	14872	53,00

Voici la signification des variables-caractères :

poptop est la population totale,
popetr est la population étrangère,
logsoc est le parc de logement social,
chomage est le taux de chomage (d'après l'INSEE),
txhab est le montant de la taxe d'habitation,
revenu est le revenu annuel moyen par habitant (sur la base des revenus en 1998),
votants est le nombre de votants au premier tour,
maj est le pourcentage obtenu par la liste majoritaire.

	poptot	popetr	logsoc	chomage	txhab	revenu	votants	maj
moyenne	207059, 66	15676,55	24,35	11,99	14,52	57436,5	59414,33	53,74
écart-type	181881,25	14549,49	13,31	5,81	4,85	36836,59	51521,14	7,44

Demi-matrice des covariances (à compléter par symétrie) :

$$\begin{bmatrix} 1 & & & & & & & \\ 0{,}89 & 1 & & & & & & \\ -0{,}28 & -0{,}24 & 1 & & & & & \\ 0{,}21 & 0{,}09 & 0{,}44 & 1 & & & & \\ 0{,}42 & 0{,}25 & 0{,}15 & 0{,}14 & 1 & & & \\ -0{,}22 & -0{,}14 & -0{,}58 & -0{,}39 & -0{,}66 & 1 & & \\ 0{,}99 & 0{,}87 & -0{,}32 & 0{,}20 & 0{,}43 & -0{,}21 & 1 & \\ -0{,}32 & -0{,}31 & -0{,}41 & -0{,}25 & -0{,}53 & 0{,}79 & -0{,}31 & 1 \end{bmatrix}$$

valeurs propres	3,59	2,40	0,95	0,53	0,30	0,13	0,10	0,00
poids	44,89	29,98	11,84	6,63	3,77	1,62	1,20	0,06

Voici les coordonnées des projections des villes dans le premier plan factoriel et les qualités de représentation.

	coordonnée 1	coordonnée 2	qualité
Poitiers	0,31	1,03	0,27
Reims	0,30	1,16	0,42
Marseille	−5,02	−2,18	0,97
Le Havre	−0,09	0,79	0,34
Besançon	0,43	0,65	0,27
Strasbourg	−0,73	−0,63	0,31
Nice	−1,93	−0,91	0,71
Lyon	−1,84	−1,53	0,81
Amiens	−0,01	1,34	0,84
St Denis Réunion	−0,16	2,15	0,38
Nancy	0,91	0,34	0,46
Toulouse	−1,65	−1,03	0,91
Brest	0,70	0,18	0,30
Boulogne Bill.	2,20	−1,68	0,95
Evry	0,24	2,32	0,74
Versailles	1,62	−0,37	0,56
Neuilly	4,44	−3,47	0,94
St Denis	0,29	1,83	0,41

Index d'Algèbre

Index d'Analyse

Crédit Photographique

p. 435 : copyright Serge Savoyski.

49629 - (II) - (1,5) - OSB 80° - AUT - MLN

Achevé d'imprimer sur les presses de
SNEL Grafics sa
Z.I. des Hauts-Sarts - Zone 3
Rue Fond des Fourches 21 – B-4041 Vottem (Herstal)
Tél +32(0)4 344 65 60 - Fax +32(0)4 289 99 61
septembre 2006 — 38841

Dépôt légal : mars 2006, suite du tirage : septembre 2006

Imprimé en Belgique



François Liret

MATHS EN PRATIQUE

À l'usage des étudiants

Ce manuel propose un cours complet de mathématiques appliquées. Il est l'outil indispensable des étudiants.

Ne supposant aucun prérequis, l'ouvrage présente les méthodes de raisonnement et d'analyse mathématiques, les outils de calcul ainsi que de nombreux exemples de modélisation dans différents domaines (physique, biologie, économie...).

Organisé en deux parties, Algèbre et Analyse, le cours est complété par des applications concrètes corrigées.

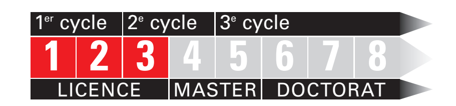
Des exercices en fin de chapitre proposent par ailleurs au lecteur de mettre en pratique les techniques opératoires du cours. Leurs corrigés sont disponibles en ligne sur le site www.dunod.com.

FRANÇOIS LIRET

est maître de conférence à l'université Paris 7 – Denis Diderot.









ISBN 2 10 049629 8

www.dunod.com